One-Shot Object Segmentation

Jagpreet Chawla, Amit Kumar Yadav, Bivas Maiti
School of Informatics, Computing & Engineering
Indiana University, Bloomington

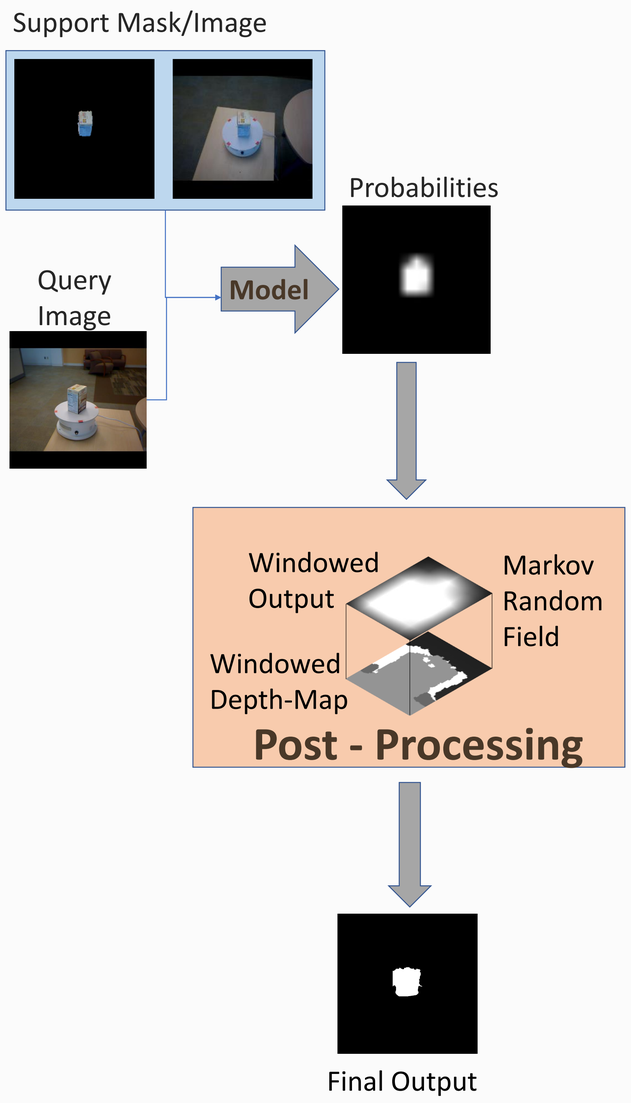
ABSTRACT

Semantic segmentation is a classic vision problem which has many applications which include, but is not limited to, robotics, autonomous driving, virtual reality, etc. Unlike detection, segmentation provides more precise object boundary which is especially useful in the field of robotics. Semantic segmentation is a highly researched topic, and most of the proposed methods require a large amount of data to train accurate models. While getting image data is not very difficult because of the internet, getting a large amount of annotated data requires a lot of resources. Therefore, in this project we are focusing on the problem on one-shot object segmentation, i.e. Segmenting an object using only one annotated example per class. We implemented approaches from two papers, compared and evaluated them on RGB-D Object Dataset, and improved the results by incorporating depth information using MRF.

Introduction

- Semantic Segmentation is per-pixel association of an image with different classes.
- This is a data intensive task and requires large amount of annotated data which is difficult and expensive to obtain.
- This is the motivation for **One-Shot Semantic Segmentation**, which requires only **1 image per class**.
- We are implementing two existing one-shot semantic segmentation models but on a different dataset.
  - **OSLSM[1]** - uses a meta-learner approach
  - **Revolver[2]** - uses guided networks.
- We then incorporated depth-map at post-processing step to improve results.

Experimental Setup

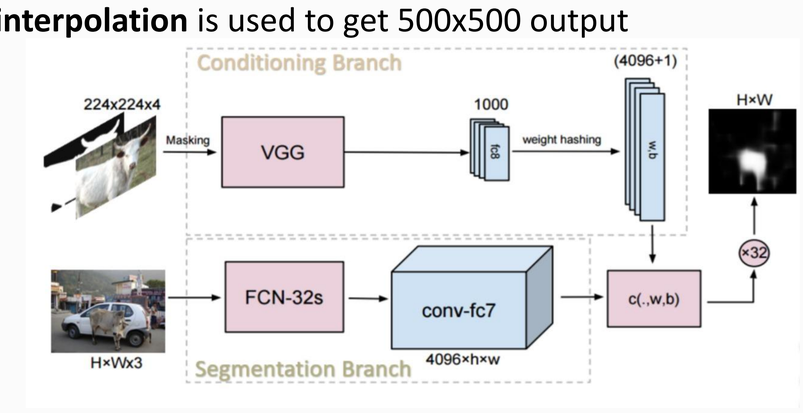


Dataset

- We are using **RGB-D object dataset [3]** that contains images of **300** common household objects.
- For each image, we also have a depth map available which was obtained using a **Kinect style 3D camera**.
- The dataset contains images from multiple frames of several short video clips.
- We are using only **4 frames** from each of the scene. We are then generating unique pairs as our dataset.
- **2 objects classes** and few scenes are kept for testing. Total **3408** image pairs are used for training.

OSLSM (One-Shot Learning for Semantic Segmentation)[1]

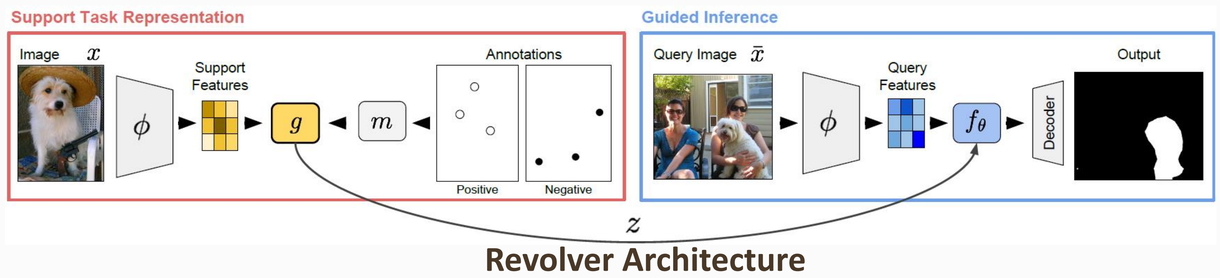
- This model takes a **meta-learner** approach.
- The model is divided into **parallel branches**:
  - **Feature Extractor**: Extracts features from query image.
  - **Meta-learner**: Learns to predict weights for a logistic regression model using masked support image. A logistic regression model with these weights is applied to extracted features.
- A predefined fixed weight matrix is used as a **hashing method** to convert **1000** dimensional VGG16 output into **4097 dimension.**
- A **pixel level logistic regression** is applied to **16x16x4096** output of **FCN32**'s last convolution layer to get **16x16 segmentation output**
- **Bilinear interpolation** is used to get 500x500 output



OSLSM Architecture
Reprinted from [1]

Revolver[2]

- Revolver has 2 branches and uses a **guided network** approach.
- Guide branch generates useful latent representations from support.
- This is the first few-shot segmentator that can work with **sparse annotations** for support images.
- Revolver extends the fully convolutional network architecture and can be trained **end-to-end**.
- Easy and fast to train and quickly updates given more guidance.



Revolver Architecture
Reprinted from [2]

Incorporating Depth(Post-Processing)

- We improved the results of above models by incorporating depth information.
- **Markov Random Field** using **graph cut method** was used for refining the results.

Results:

Model Performance (mean Intersection over union)

(On positive samples)

	Unseen classes		Seen classes	
Model	RGB	RGB-D	RGB	RGB-D
OSLSM	32.58%	42.64%	51.11%	63.73%
Revolver	14.76%	15.39%	18.76%	22.36%

Contribution

- Implemented two papers, but on a **different dataset**.
- Used a weighted log loss to account for the imbalance of positive and negative pixels. Below is the loss function:
  $-((w_0)(y)\log(p)+(w_1)(1-y)\log(1-p))$
  where w_0 and w_1 are weights calculated for positive and negative segments.
- Incorporated **depth channel** as a post processing step to improve the results.

Discussion

- OSLSM performs well on positive samples from dataset of simple images.
- Revolver has a blocky output which we are trying to mitigate by trying out different loss functions in consultation with the original author.
- Reason behind this can be the smaller size of the classes in this dataset.
- OSLSM is able to perform well even on **unseen** classes.
- OSLSM is not performing well on negative examples, i.e. it's sometimes segmenting out wrong objects.
- To mitigate this issue, in future we plan to train network with more negative samples and impose high penalty if it segments wrong object.

Acknowledgement

We would like to thank Dr. Md Alimoor Reza for his invaluable guidance in this project.

References

[1] Amirreza Shaban et al. "One-Shot Learning for Semantic Segmentation". In: arXiv e-prints, arXiv:1709.03410 (Sept. 2017), arXiv:1709.03410. arXiv: 1709.03410 [cs.CV].

[2] Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., & Levine, S. (2018). Conditional networks for few-shot semantic segmentation.

[3] Lai, Kevin, et al. "A large-scale hierarchical multi-view rgb-d object dataset." 2011 IEEE international conference on robotics and automation. IEEE, 2011.