Unsupervised Cross Domain Semantic Segmentation

Shanu Kumar (150659)
Department of Electrical Engineering
sshanu@iitk.ac.in

Sourabh Kumar (170716)
Department of Computer Science and Engineering
ssourabh@iitk.ac.in

1 Introduction

Semantic segmentation aims to assign each pixel a semantic label, e.g., person, car, road or tree, in an image. Semantic Segmentation on the dataset provided can be useful for security purposes, as we can segment only the cars, motorbikes, persons, and cycles. Therefore, we aim to do semantic segmentation on the provided dataset using two approaches. The first approach is based on the domain adaptation of semantic segmentation in an unsupervised manner, and the last method tries to solve the task of segmentation in the unsupervised setup. As both the approaches are unsupervised, these can be applicable on the provided dataset.

2 Methods

2.1 Learning to Adapt Structured Output Space for Semantic Segmentation

In unsupervised domain adaptation problem, there are two set of images from source and target domains, which are denoted as $\{I_s\}$ and $\{I_t\}$. In source domain, images have their segmentation ground truth annotations denoted as $\mathbf{Y}_s$. There are many approaches based on feature adaptation for solving the task of image classification in this setup. However, feature adaptation for semantic segmentation may suffer due to high dimensional size of features which are needed to encode diverse visual cues like shape, appearance, and context. Tsai *et al.* [2018] propose a method based on adversarial learning in the output space, as the output space are low dimensional and con

The proposed model in Tsai *et al.* [2018] consists of two modules: 1) a segmentation network $\mathbf{G}$ to output the segmentation of the input image and 2) a discriminator to distinguish whether the input image is from source or target segmentation output. The overview of the model is shown in the figure 1 Due to adversarial training, the proposed segmentation network $\mathbf{G}$ will fool the discriminator and generate output from the same distribution for both source and target images.

2.1.1 Segmentation Network

For segmentation, we follow the same approach as proposed in Tsai *et al.* [2018]. DeepLab-v2 framework Chen *et al.* [2018] is adopted as the baseline model with ResNet-101 model pre-trained on ImageNetDeng *et al.* [2009].
Segmentation loss is defined using the cross-entropy loss function for images from the source images:

$$\mathscr{L}_{seg}(I_s) = -\sum_{(h,w)}\sum_{c\in C} Y_s^{(h,w,c)} \log({P_s}^{(h,w,c)})$$

where Y_s is the ground truth annotations for source images and $P_s = \mathbf{G}(I_s)$ is the segmentation output.

2.1.2 Discriminator

The architecture of the discriminator proposed in Tsai *et al.* [2018] consists of 5 convolution layers with kernel (4×4) and stride of 2, where the channel number is $\{64, 128, 256, 512, 1\}$ respectively.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

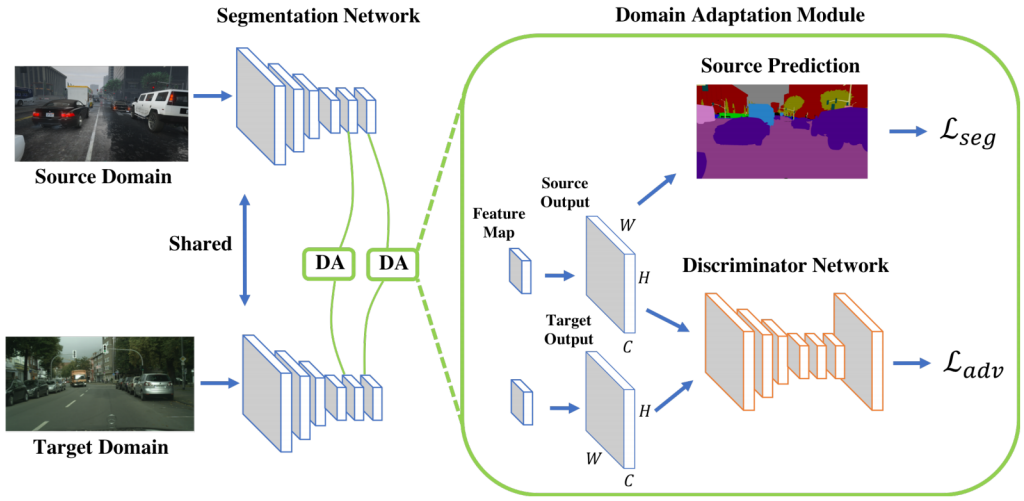

Figure 1: Model overview of Tsai *et al.* [2018]

For training the discriminator, segmentation softmax output $P = \mathbf{G}(I)$ is forwarded to the discriminator, and it is trained using a cross-entropy loss L_d for the two classes (i.e., source and target). The loss is defined as:

$$\mathscr{L}_d(P) = -\sum_{(h,w)} (1-z)\log(\mathbf{D}(P)^{(h,w,0)}) + z\log(\mathbf{D}(P)^{(h,w,1)})$$

where z = 0 if the sample is drawn from the target domain, and z = 1 for the sample from the source domain.

2.1.3 Multi-level Adversarial Learning

Due to adaptation in the output space, features will also get adapted as the gradients are back-propagated. But there is one issue in the proposed model, the low-level features may not be adapted well as they are far away from the output level. To fix this issue, they propose to incorporate adversarial learning at different feature levels of the segmentation network. Here, Tsai *et al.* [2018] have used $conv5$ and $conv4$ features to predict segmentation outputs in the output space. Hence the adversarial loss $\mathscr{L}^i_{adv}$, where i indicates the level is defined as:

$$\mathscr{L}^i_{adv}(I_t) = -\sum_{(h,w)} \log(\mathbf{D}(P_t)^{(h,w,1)})$$

This loss is used to train the segmentation network to fool the discriminator by maximizing the probability of the image from the target domain to be considered as to be from the source domain.

The final objective function is defined as:

$$\mathscr{L}(I_s, I_t) = \sum_i \lambda^i_{\text{seg}} \mathscr{L}^i_{seg}(I_s) + \sum_i \lambda^i_{\text{adv}} \mathscr{L}^i_{adv}(I_t)$$

where i indicates the level used for predicting the segmentation output.

The model is trained by optimizing the following the criterions:

$$\max_{\mathbf{D}} \min_{\mathbf{G}} \mathscr{L}(I_s, I_t)$$

2.2 Unsupervised Image Segmentation by Backpropagation

In Kanezaki [2018], propose a method for solving the segmentation task in an unsupervised setup. They first assing cluster labels to all the pixels of the image by $c_n = f(\mathbf{x_n}$, where f is a assignment

function which will return cluster centriod closet to $\mathbf{x}_n$ among k centriods, which are obtained by e.g. k-means clustering. They propose to solve two problems, 1) when f and feature representation $\{\mathbf{x}_n\}$ are fixed, then $\{c_n\}$ are obtained using the above method 2) when f and $\{\mathbf{x}_n\}$ are trainable, but $\{c_n\}$ are fixed or given, then this is a standard supervised classification problem. Hence they solve both the problems alternatively: predicting the optimal $\{c_n\}$ with fixed f and $\{\mathbf{x}_n\}$, and training of the parameters of f and $\{\mathbf{x}_n\}$ with fixed $\{c_n\}$. The proposed model in Kanezaki [2018] is shown in the figure 2.

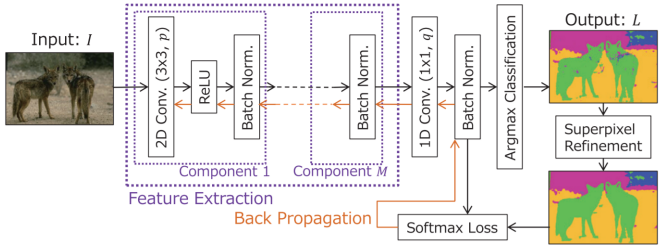

Figure 2: Model overview of Kanezaki [2018]

3 Experiments

We have experimented with domains and pre-training of the model proposed in Tsai *et al.* [2018] in to achieve better results. We have used Cityscapes Cordts *et al.* [2016] and GTA5 Richter *et al.* [2016] as source or target domains in these experiments.

GTA5 → Cityscapes: We have used the available model on the github repository of the paper Tsai *et al.* [2018], which is already adapted from GTA5 to Cityscapes.

GTA5 → IITK: We trained the proposed model in Tsai *et al.* [2018] with the same set of hyper-parameters, but source domain as GTA5 and target domain as the dataset given (IITK) with two different settings. In first setting, we initialize the model with adapted model (GTA5 → Cityscapes). In the other setting, we train the model from scratch without any pre-training.

Cityscapes → IITK: We trained the proposed model in Tsai *et al.* [2018] with the same set of hyper-parameters, but source domain as Cityscapes and target domain as IITK. Here, we initialize the model with adapted model (GTA5 → Cityscapes), because the number of images in the Cityscapes is 2975, whereas in GTA5 it is 24966. Hence here we cannot train the model from scratch.

GTA5 → Cityscapes + IITK: We trained the proposed model in Tsai *et al.* [2018] with the same set of hyper-parameters, but source domain as GTA5 and target domain is set as dataset combined frmo IITK and Cityscapes. In this case, we also train the model without pre-training.

4 Results

From the figure 3, we can observe that both the models which are adapted from GTA5 to IITK, perform better than the model which is adapted from GTA5 to Cityscapes. This shows the significance of adversarial adaptation of the features in the output space. The model which is trained from the scratch and adapted from GTA5 to IITK, is not optimized well due to limited computing resources and time, but still it performs well.

From the figure 4, we observe that the model adapted from Cityscapes to IITK perform better than the models adapted from GTA5 to IITK. The reason is that the IITK dataset is more similar to the Cityscapes dataset than the GTA5 dataset. Therefore, when the model which is already adapted from GTA5 to Cityscapes, is trained on Cityscapes and adapted to IITK will perform better. In the last column of the figure 4, we observe decent results, as the model is trained from scratch or without any pre-training, hence due to limited resources we are not able to optimize the model well.

We also compared the results from the model Kanezaki [2018], but the figure 4 suggests that the segmentation outputs are rich but the we want to segment only the selected objects, which is not the case here.

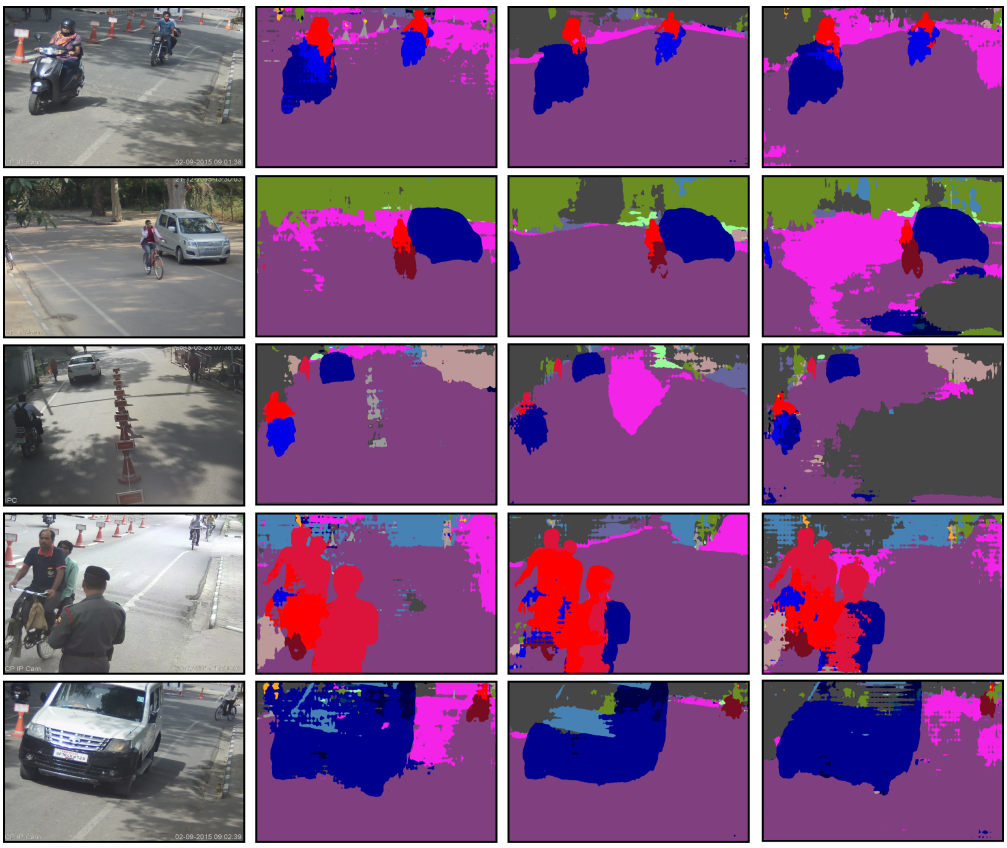

Figure 3: Visualization of the segmentation results from different models. The first column shows the input image. The second column represents the segmentation outputs from the already adapted model from GTA5 to Cityscapes. The third column shows the results from model which is adapted from GTA5 to IITK but the model is initialized with the adapted model from GTA5 to Cityscapes. The last columnn represents the results from the model trained from GTA5 to IITK but without any pre-training or initialization.

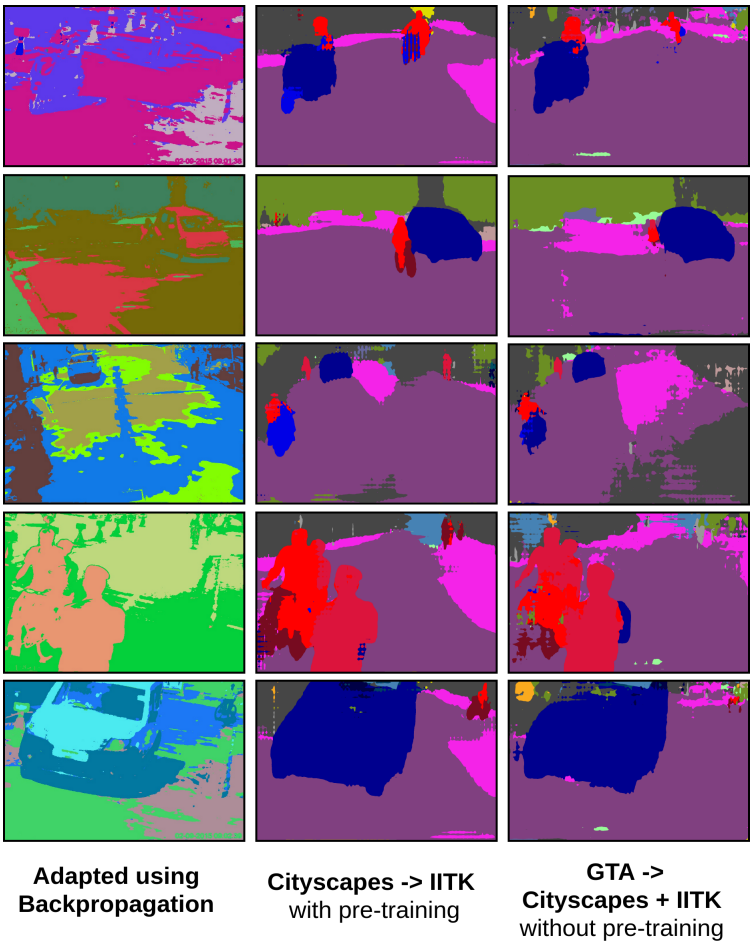

Figure 4: Visualization of the segmentation results from different models. The first column shows the segmentation output using the approach proposed in Kanezaki [2018]. The second column shows the results from the adaptation of the pre-trained model from Cityscapes to IITK dataset. The last column represents the segmentation results using the model which is adapted from GTA5 to Cityscapes and IITK.

References

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Asako Kanezaki. Unsupervised image segmentation by backpropagation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1543–1547. IEEE, 2018.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016.

Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018.