Domain Adaptation for Image Segmentation

Introduction and Motivation

- Annotating real data for image segmentation is laborious and time consuming task.
- Annotation of single frame of Cityscapes dataset takes about 1 hr.
- We aim to adapt the representation learned on synthetic data to real world data.

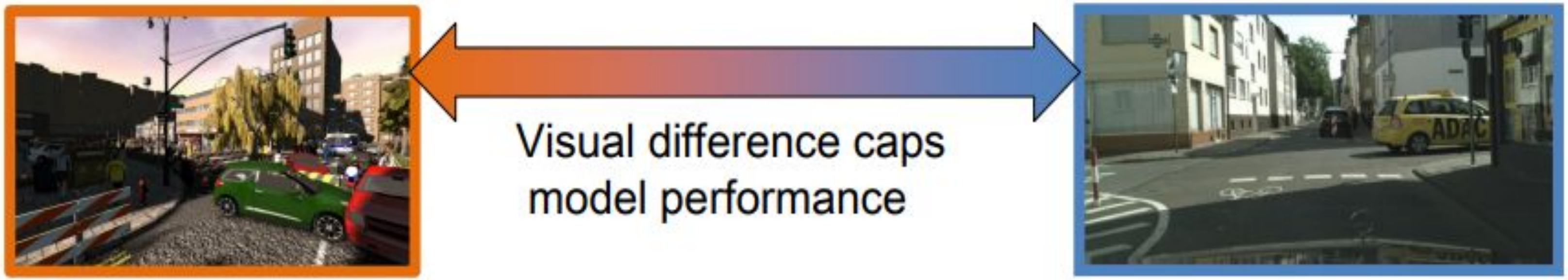

Image Source : https://goo.gl/12iVHF

Related Work

1. Exploiting the power of CNN, *Shelhamer et al* proposed FCN for semantic pixel labelling task but failing to address the challenge of domain shift within the context of semantic segmentation.
2. For deep domain adaptation some approaches use Maximum Mean Discrepancy while some use adversarial approaches.
3. Many approaches like PixelDA and CoGAN techniques perform adaptation for classification task on pixel space.
4. For semantic segmentation *Hoffman's* FCN in the wild and *Zhang* curriculum domain adaptation addresses the problem.
5. One concurrent work CyCADA uses CycleGAN and transfers the source domain images to target domain with pixel alignment.

Related Work

- Long et al. proposed CNN models can be converted to fully-convolutional network for semantic segmentation.
    - Difficult to obtain annotations
    - May not generalize well to unseen image domains.
- Hoffman et al. introduced Domain adaptation by applying adversarial learning in a fully-convolutional way on feature representations.
- CyCADA transfers source domain images to the target domain with pixel alignment.
    - Generates extra training data along with feature space adversarial learning.

Datasets

Source:

- **GTA-5:** It is curated from the frames of a popular game, ***Grand Theft Auto V*** and consists of 24,966 densely labelled frames.
- **Synthia:** 9400 frames with semantic annotation compatible with Cityscapes.

Target:

- **Cityscapes:** Urban street images of 50 cities with 5000 images.
- ***Indian Roads:*** Videos obtained from youtube.

Proposed Model

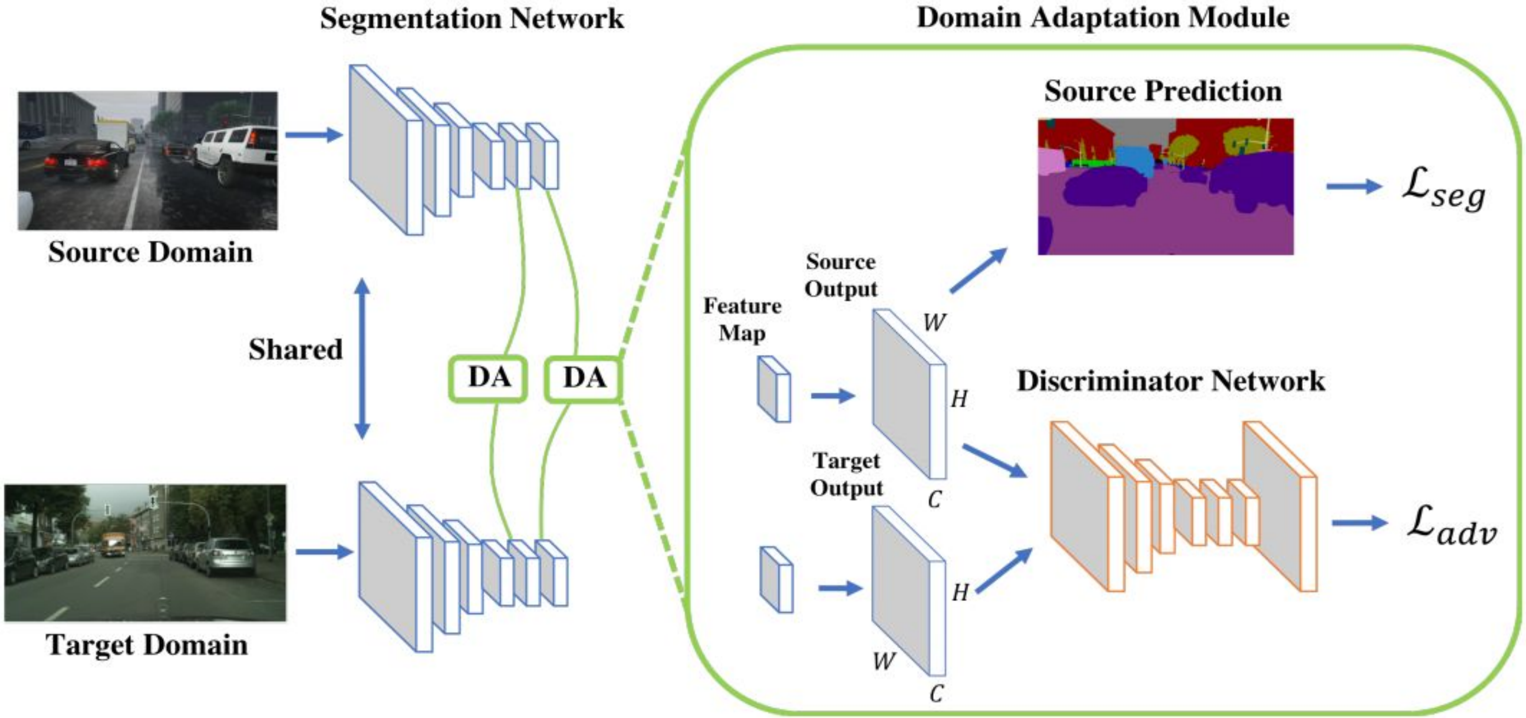

Image Source : https://goo.gl/EcBAq3

Proposed Model

- **Generator G (Segmentation Network)**:
    - Source image is forwarded through the segmentation network to predict the segmentation softmax output.
    - Adversarial loss on the target prediction makes the G generate similar segmentation distribution in the target domain to the source prediction.

- **Discriminator D:**
    - Discriminator is trained to distinguish between the source and target domain.

Objective Function

- **Segmentation Loss:** Cross-entropy loss using the ground truth annotations in the source domain.
- **Adversarial Loss:** Helps target predictions to adapt to the distribution of the source predictions.
- **Discriminator Loss:** Cross-entropy loss for the two classes (i.e., source and target)

Network Architecture

- **Discriminator:**
  - 5 CONV layers with 4x4 kernel and stride of 2 (no. of output channels: (64, 128, 256, 512, 1).
  - Except for last CONV layer, each layer is followed by a leaky ReLU.
  - Last layer is followed by a upsampling layer to rescale the output to the size of input.
- **Segmentation Network:**
  - A deep convolutional net (VGG-16) with transformed fully connected layers to convolutional layers.
  - Modified stride of last 2 convolutional layers from 2 to 1.
  - An up-sampling layer along with the softmax output to match the size of the input image.

Evaluation

- We pick the common classes between the source and target and evaluate in terms of IOU of these classes.

For instance,

- ***GTA-5 -> Cityscapes*** contain 19 similar classes.
- ***Synthia -> Cityscapes*** contain 16 similar classes.

We compute mean IOU of the similar classes as the metric.

References

- https://arxiv.org/pdf/1802.10349.pdf
- https://arxiv.org/pdf/1711.06969.pdf
- http://synthia-dataset.net/
- https://download.visinf.tu-darmstadt.de/data/from_games/
- https://www.cityscapes-dataset.com/

Thank You

Anshul Khantwal
Deepak Thukral
Sharat Agarwal