Unsupervised Domain Adaptation with Multiple Domain Discriminators and Adaptive Self-Training

Teo Spadotto, Marco Toldo, Umberto Michieli, Pietro Zanuttigh

January 15th, 2021
Semantic Segmentation

- Dense labeling task: assign a class label to each single pixel in an image
- Nowadays solved with deep learning, typically auto-encoder CNNs
- Large generic datasets for training are available but it is challenging to get data specific to the task
Unsupervised Domain Adaptation

- Labeling data is available only for the source dataset
- Goal: achieve good results on a different (but related) target dataset
- Domain shift limits performance: need for adaptation
- The adaptation can be performed at input, feature or output space
Output Level Adaptation

**Task Loss:** extract knowledge from source supervision

\[ \mathcal{L}_{full} = \mathcal{L}_{G,0} + w_1^{s,t} \mathcal{L}_{G,1}^{s,t} + w_2^t \mathcal{L}_{G,2}^t + w_3 \mathcal{L}_{G,3} \]

**Self-Training:** use confident target pseudo labels for self-taught supervision

**Output Level Adversarial Modules:** align prediction maps between source and target domains w.r.t. source GT maps

**Examples:**
- Synth. GT
- Synth. RGB
- Real RGB
- Real segmentation
- Synth. segmentation

**Adaptive Thresholding**

\[ T_f[c, X_n] \]
Adversarial Adaptation

Two Adversarial Adaptations

- **D1: GT vs Prediction**
  - ⇒ indirect domain alignment
  - ⇒ both source and target predictions can be used
- **D2: Source vs Target**
  - ⇒ direct domain alignment

\[
\mathcal{L}_{G, 1}^{s,t} = - \sum_{p \in X_n^{s,t}} \log(D_1(G(X_n^{s,t}))^{(p)})
\]
\[
\mathcal{L}_{D_1} = - \sum_{p \in X_n^{s,t}} \log(1 - D_1(G(X_n^{s,t}))^{(p)}) + \log(D_1(Y_n^{s})^{(p)})
\]
\[
\mathcal{L}_{G, 2}^{t} = - \sum_{p \in X_n^{t}} \log(D_2(G(X_n^{t}))^{(p)})
\]
\[
\mathcal{L}_{D_2} = - \sum_{p \in X_n^{t}} \log(1 - D_2(G(X_n^{t}))^{(p)}) + \log(D_2(Y_n^{t})^{(p)})
\]
Self-Training

Loss: Weighted cross-entropy with pseudo labels

\[ \mathcal{L}_{G,3} = -\sum_{p \in X_n^t} \sum_{c \in C} \mathcal{M}_f^{(p)} \cdot W_c \cdot \hat{Y}_n^{(p)}[c] \cdot \log(G(X_n^t)^{(p)}[c]) \]

- Use highly confident network predictions for self-teaching on target dataset
- Use discriminator’s output as a confidence measure
- Class and step adaptive thresholding dynamically updated during training

Confidence based mask computed from discriminator’s output

Adaptive Threshold based on class-wise distribution
### Quantitative Results

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GTA5→CS</td>
</tr>
<tr>
<td>Supervised (baseline)</td>
<td>31.9</td>
</tr>
<tr>
<td>Hoffman et al. [1]</td>
<td>27.1</td>
</tr>
<tr>
<td>Hung et al. [2]</td>
<td>29.0</td>
</tr>
<tr>
<td>Zhang et al. [3]</td>
<td>28.9</td>
</tr>
<tr>
<td>Biasetton et al. [4]</td>
<td>30.4</td>
</tr>
<tr>
<td>Michieli et al. [5]</td>
<td>33.3</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>35.1</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GTA5→MAP</td>
</tr>
<tr>
<td>Supervised (baseline)</td>
<td>37.8</td>
</tr>
<tr>
<td>Hung et al. [2]</td>
<td>34.4</td>
</tr>
<tr>
<td>Biasetton et al. [4]</td>
<td>35.2</td>
</tr>
<tr>
<td>Michieli et al. [5]</td>
<td>38.5</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>41.9</strong></td>
</tr>
</tbody>
</table>

---

- 2 Source synthetic datasets (GTA5 or SYNTHIA)
- 2 Target real-world datasets (Cityscapes and Mapillary)
- Results computed using a DeepLab-v2 network with Resnet-101 as encoder

---

Visual Results (Cityscapes)
### Visual Results (Mapillary)

<table>
<thead>
<tr>
<th>road</th>
<th>sidewalk</th>
<th>building</th>
<th>wall</th>
<th>fence</th>
<th>pole</th>
<th>traffic light</th>
<th>traffic sign</th>
<th>vegetation</th>
<th>terrain</th>
<th>sky</th>
<th>person</th>
<th>rider</th>
<th>car</th>
<th>truck</th>
<th>bus</th>
<th>train</th>
<th>motorcycle</th>
<th>bicycle</th>
<th>unlabeled</th>
</tr>
</thead>
</table>

#### From GTA5
- Image
- Annotation
- Supervised ($\mathcal{L}_{G,1}$)

#### To Mapillary
- Image
- Annotation
- Bialetta et al. [4]
- Ours ($\mathcal{L}_{full}$)
Conclusions

- We presented a novel adversarial learning and self-teaching scheme for unsupervised domain adaptation.
- Domain discriminators capture both source vs target and ground truth vs prediction statistics.
- Adaptive self-training strategy.
- Experimental results on synthetic to real adaptation show that the approach outperforms competing schemes using output-level adaptation.

Paper webpage: [https://lttm.dei.unipd.it/paper_data/semanticDA](https://lttm.dei.unipd.it/paper_data/semanticDA)