Adaptive Deep Learning for Visual Understanding

Kate Saenko
Boston University
AIR: AI Research at BU
Goal: teach machines to see, talk, act

- A baseball game in progress with the batter up to plate
- A man is riding a bicycle
- Q: What is the child standing on?
- A: skateboard
- Find “window upper right”
- Find the moment when “girl looks up at the camera and smiles”
- “Go out of the bedroom, down the stairs, turn left and stop at the dining table”

Vision

Language

Action
deep learning

human learning
adaptive
explainable
modular
adaptive, explainable, modular

HUMAN
- learns from a single example
- generalizes knowledge
adaptive, explainable, modular

SUPERVISED DEEP LEARNING

- learns from 10000’s examples
- fails on new domains
adaptive, **explainable**, modular

tail, four legs, ...

why is this a dog?

HUMAN

- can explain decisions
- grounds language in world
DEEP LEARNING

- cannot explain decisions
- black box

why is this a dog?

adaptive, explainable, modular
adaptive, explainable, modular

HUMAN

- disentangles properties
- combinatorial
adaptive, explainable, modular

DEEP LEARNING

- entangles properties
- not compositional
this talk

adaptive
explainable
modular
Domain shift

What your net is trained on

What it’s asked to label

“Dataset Bias”
“Domain Shift”
Example shift

from simulation

train

test

to reality
Shift from simulation to reality

Input Image

True Segmentation

Model Output
Solution: Domain Adaptation

Input Image

True Segmentation

trees
car
people
sidewalk
road

Adapted Model Output

Model Output
glass
sidewalk
car
Applications of Domain Adaptation

- From dataset to dataset
- From RGB to depth
- From simulated to real control
- From CAD models to real images
Adversarial domain adaptation

Source Data + Labels

Encoder CNN

Classifier

Classifier loss
Adversarial domain adaptation

Source Data + Labels
- backpack
- chair

Unlabeled Target Data

Encoder CNN → Classifier

Adversarial domain adaptation

Encoder CNN

Classifier
Adversarial domain adaptation

Source Data + Labels
- backpack
- chair

Unlabeled Target Data

Encoder CNN

Classifier

Discriminator

Adversarial domain adaptation

Classifier loss

Adversarial loss
Adversarial domain adaptation

- **Source Data + Labels**
  - backpack
  - chair

- **Unlabeled Target Data**

- **Encoder CNN**

- **Classifier**
  - Classifier loss

- **Discriminator**
  - Adversarial loss

- Adversarial domain adaptation
Results on Digits Classification

- Domain Confusion Loss (Tzeng 2015)
- Adversarial Domain Alignment (ADDA) (Tzeng 2017)

<table>
<thead>
<tr>
<th>METHOD</th>
<th>SVHN to MNIST</th>
<th>USPS to MNIST</th>
<th>MNIST to USPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Only</td>
<td>67.1</td>
<td>68.1</td>
<td>77.0</td>
</tr>
<tr>
<td>GradReversal (Ganin &amp; Lempitsky 2015)</td>
<td>73.9</td>
<td>73.0±2.0</td>
<td>77.1±1.8</td>
</tr>
<tr>
<td>Domain Confusion Loss (Tzeng 2015)</td>
<td>68.1±0.3</td>
<td>66.5±3.3</td>
<td>79.1±0.5</td>
</tr>
<tr>
<td>ADDA (Tzeng 2017)</td>
<td>76.0±1.8</td>
<td>90.1±0.8</td>
<td>89.4±0.2</td>
</tr>
</tbody>
</table>

Eric Tzeng, Judy Hoffman, Trevor Darrell, Kate Saenko. “Simultaneous Deep Transfer Across Domains and Tasks” ICCV 2015
Problem: ambiguous features
Problem: ambiguous features

Source Data + Labels

Encoder CNN

Unlabeled Target Data

Encoder CNN

Classifier

Classifier loss
Problem: ambiguous features

Source Data + Labels
- backpack
- chair

Unlabeled Target Data

Encoder CNN

Classifier
- Classifier loss

Discriminator
- Adversarial loss
Problem: ambiguous features

Source Data + Labels

Unlabeled Target Data

Encoder CNN

Classifier

Discriminator

Classifier loss

Adversarial loss

backpack

chair

Unlabeled Target Data

Kate Saenko, Boston University
Problem: ambiguous features

Before Adaptation

Adapted

Source

Target
Goal: avoid generating ambiguous features
Solution: use the decision boundary

Train a critic ($C$) (=discriminator) that can detect target samples near decision boundary

Train a generator ($G$) that can fool the critic

Slightly change the boundary and measure the change of $p(y|x)$! (sensitivity)

Samples near the boundary have larger sensitivity
Adversarial Dropout Regularization

Data $\rightarrow G \rightarrow G(x) \rightarrow C_1, C_2$ 

Fix $G$ and train $C$ to maximize $d(p_1, p_2)$ for target samples.
Train $G$ and $C$ to minimize Cross Entropy for source samples.
For $k = 1:n$
  Fix $C$ and train $G$ to minimize $d(p_1, p_2)$ for target.
Results on Digits Classification

- Adversarial Dropout Regularization (ADR) (Saito et al 2018)

<table>
<thead>
<tr>
<th>METHOD</th>
<th>SVHN to MNIST</th>
<th>USPS to MNIST</th>
<th>MNIST to USPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Only</td>
<td>67.1</td>
<td>68.1</td>
<td>77.0</td>
</tr>
<tr>
<td>ATDA (Saito et al. (2017))</td>
<td>86.2†</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GradReversal (Ganin &amp; Lempitsky 2015)</td>
<td>73.9</td>
<td>73.0±2.0</td>
<td>77.1±1.8</td>
</tr>
<tr>
<td>Domain Confusion Loss (Tzeng 2015)</td>
<td>68.1±0.3</td>
<td>66.5±3.3</td>
<td>79.1±0.5</td>
</tr>
<tr>
<td>ADDA (Tzeng 2017)</td>
<td>76.0±1.8</td>
<td>90.1±0.8</td>
<td>89.4±0.2</td>
</tr>
<tr>
<td>CoGAN (Liu &amp; Tuzel (2016))</td>
<td>did not converge</td>
<td>89.1±0.8</td>
<td>91.2±0.8</td>
</tr>
<tr>
<td>DTN (Taigman et al. (2016))</td>
<td>84.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>96.7±1.85</td>
<td>91.5±3.61</td>
<td>91.3±0.65</td>
</tr>
</tbody>
</table>

Adversarial Dropout Regularization, Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, Kate Saenko, ICLR 2018
Adaptation for semantic segmentation
Adaptation for semantic segmentation

Input image

Ground Truth

Adapted

Source-only model
Pixel-to-pixel adaptation

Source only

Source: SVHN

Target Accuracy: 62.3%

→

CyCADA

Target: MNIST

Target Accuracy: 86.6%
Strong-weak feature alignment

Source  
Target

Global-Weak Alignment

Strong-Weak Distribution Alignment for Adaptive Object Detection, Saito, Ushiku, Harada, Saenko, arxiv 2018
Strong-weak feature alignment

Strong-Weak Distribution Alignment for Adaptive Object Detection, Saito, Ushiku, Harada, Saenko, arxiv 2018
Domain shift: from PASCAL VOC to CLIPART

ours

baseline
Domain shift: from Grand Theft Auto game to Cityscapes

<table>
<thead>
<tr>
<th>Method</th>
<th>G</th>
<th>I</th>
<th>CTX</th>
<th>L</th>
<th>P</th>
<th>AP on Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster RCNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34.6</td>
</tr>
<tr>
<td>BDC-Faster</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31.8</td>
</tr>
<tr>
<td>DA-Faster</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>34.2</td>
</tr>
<tr>
<td>Weak Align</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>35.8</td>
</tr>
</tbody>
</table>

**Proposed Method with different parameters**

<table>
<thead>
<tr>
<th>Method</th>
<th>G</th>
<th>I</th>
<th>CTX</th>
<th>L</th>
<th>P</th>
<th>AP on Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFL</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>38.7</td>
</tr>
<tr>
<td>FL ($\gamma = 3$)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>42.3</td>
</tr>
<tr>
<td>FL ($\gamma = 3$)*</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>47.7</td>
</tr>
<tr>
<td>Oracle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>53.1</td>
</tr>
</tbody>
</table>

Evidence for target domain label (GradCAM) shows that the feature extractor seems to deceive the domain classifier in regions with car.
Domain shift: from Cityscapes to Foggy Cityscapes

Table 4. AP on adaptation from Cityscapes to Foggy Cityscapes (%). The performance of our method is very near to oracle, which is trained on labeled target images.

<table>
<thead>
<tr>
<th>Method</th>
<th>G</th>
<th>I</th>
<th>CTX</th>
<th>L</th>
<th>bus</th>
<th>bicycle</th>
<th>car</th>
<th>bike</th>
<th>prsn</th>
<th>rider</th>
<th>train</th>
<th>truck</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster RCNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22.3</td>
<td>26.5</td>
<td>34.3</td>
<td>15.3</td>
<td>24.1</td>
<td>33.1</td>
<td>3.0</td>
<td>4.1</td>
<td>20.3</td>
</tr>
<tr>
<td>BDC-Faster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25.0</td>
<td>31.0</td>
<td>40.5</td>
<td>22.1</td>
<td>35.3</td>
<td>20.2</td>
<td>20.0</td>
<td>27.1</td>
<td>27.6</td>
</tr>
<tr>
<td>DA-Faster</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>33.1</td>
<td>23.3</td>
<td>25.5</td>
<td>15.6</td>
<td>23.4</td>
<td>29.0</td>
<td>10.9</td>
<td>19.6</td>
<td>22.5</td>
</tr>
<tr>
<td>Proposed</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>33.5</td>
<td>33.3</td>
<td>42.7</td>
<td>22.2</td>
<td>27.1</td>
<td>40.3</td>
<td>11.6</td>
<td>22.3</td>
<td>29.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34.3</td>
<td>32.2</td>
<td>36.2</td>
<td>23.7</td>
<td>27.5</td>
<td>39.3</td>
<td>5.4</td>
<td>24.4</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>38.0</td>
<td>31.2</td>
<td>41.8</td>
<td>20.7</td>
<td>26.6</td>
<td>37.6</td>
<td>19.7</td>
<td>20.5</td>
<td>29.5</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>36.2</td>
<td>35.3</td>
<td>43.5</td>
<td>30.0</td>
<td>29.9</td>
<td>42.3</td>
<td>32.6</td>
<td>24.5</td>
<td>34.3</td>
</tr>
<tr>
<td>Oracle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50.0</td>
<td>36.2</td>
<td>49.7</td>
<td>34.7</td>
<td>33.2</td>
<td>45.9</td>
<td>37.4</td>
<td>35.6</td>
<td>40.3</td>
</tr>
</tbody>
</table>
adaptive
explainable
modular
CNNs learn to predict pneumonia by detecting hospital which took the image

- Study on detecting pneumonia using 158,323 chest radiographs
- CNNs robustly identified hospital system and department within a hospital
- CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image field of view at the time they capture the image

RISE: randomly mask input, measure output

RISE: Randomized Input Sampling for Explanation of Black-box Models, Petsiuk, Das, Saenko, BMVC 2018
input: There is a small gray block; are there any spheres to the left of it?

input image
Disentangling properties via generation

A Two-Stream Variational Adversarial Network for Video Generation

generated video
Disentangling properties via generation
Disentangling properties for domain transfer

The task is to control mouth expression in target:

with an image from source.

CycleGAN  Our Method
adaptive
explainable
modular