Detach and Adapt:
Learning Cross-Domain Disentangled Deep Representation for Image Synthesis and Classification

*Tzu-Chien Fu¹, *Yen-Cheng Liu², Wei-Chen Chiu¹,³, and Y.-C. Frank Wang¹,²

¹Research Center for IT Innovation, Academia Sinica
²Dept. EE, National Taiwan University
³Dept. CS, National Chiao Tung University

(* indicates equal contributions)
(Traditional) Machine Learning vs. Transfer Learning

• **Transfer Learning**
  • Collecting/annotating data is typically **expensive**.
  • Improved learning & understanding in the **target domain** by leveraging knowledge from the **source domain**
Research Focuses

• Transfer Learning for
  • Homogeneous/heterogeneous domain adaptation
  • Multi-label classification / zero-shot learning
  • Robust face recognition (e.g., cross-resolution, cross-modality, etc.)
Heterogeneous Domain Adaptation

• Deep Transfer Learning for Cross-Domain Data Classification
  • Learning from source & target-domain data described by distinct types of features
Heterogeneous Domain Adaptation (cont’d)

• Transfer Neural Trees (TNT)
  • Joint learning of cross-domain mapping $F_S/F_T$ & cl. layer $G$ (deep neural decision forest)
  • Propose stochastic pruning for $G$ to avoid overfitting source-domain labeled data
  • Unique embedding loss for learning target-domain data in a semi-supervised setting

Multi-Label Classification

- Predicting multiple labels w/o using annotated ground truth info (e.g., bounding box)
- Learning across image and label-domain data + exploit label co-occurrences

Labels:
- Person
- Table
- Sofa
- Chair
- TV
- Lights
- Carpet
- ...

Image of a family in a living room with a TV, sofa, and table.
Multi-Label Classification (cont’d)

• **Canonical Correlated AutoEncoder (C2AE)** [AAAI’17]
  - Unique integration of *autoencoder* & *deep canonical correlation analysis (DCCA)*
  - *Autoencoder* in C2AE: label embedding + label recovery + label co-occurrence
  - *DCCA* in C2AE: joint feature & label embedding

Y.-C. F. Wang et al., Learning Deep Latent Spaces for Multi-Label Classification, AAAI 2017
Research Focuses

• Transfer Learning for
  • Domain adaptation
    • Cross-domain image synthesis/translation/classification
  • Multi-label classification / zero-shot learning
  • Robust face recognition (e.g., cross resolution, etc.)
FaceApp

- Beyond putting a smile on your face
- Over 10M downloads
Introduction

• Feature Disentanglement:
  • Learn a latent space which factorizes the representation $z$ into different parts (i.e., attributes) for describing the corresponding info (e.g., identity, pose, or expression of facial images).
Settings for Feature Disentanglement

• **Unsupervised Learning**
  - Disentangling images *without* observing attribute info
  - *No guarantee* in disentangling particular semantics

• **Supervised Learning**
  - With supervision of image labels, disentangle the associated factor from feature representation
  - Can manipulate the output image with label/attribute of interest accordingly.

• **Ours: Cross-Domain Feature Disentanglement**
  - Source-domain training data: existing annotated instances
  - Target-domain data: no ground truth info, to be adapted/manipulated
  - Can be viewed as either *semi-supervised learning*, or *unsupervised domain adaptation*
Our Goal

• A unified framework for cross-domain feature disentanglement, with only attribute supervision from the source domain.
Related Works

• Feature Disentanglement
  • Unsupervised: InfoGAN [1]
  • Supervised: AC-GAN [2]

• Unsupervised Cross-Domain Image Synthesis/Translation
  • Image synthesis: CoGAN [3]
  • Image translation: UNIT [4]

InfoGAN & AC-GAN (Unsup/Sup. Feature disentanglement)

InfoGAN (unsupervised)

\[
\min_{G} \max_{D} V(D,G) = V(D,G) - \lambda I(c; G(z,c))
\]

AC-GAN (supervised)

\[
L_S = E[\log P(S = \text{real} \mid X_{\text{real}})] + E[\log P(S = \text{fake} \mid X_{\text{fake}})] \\
L_C = E[\log P(C = c \mid X_{\text{real}})] + E[\log P(C = c \mid X_{\text{fake}})]
\]


CoGAN (Unsupervised Cross-Domain Image Synthesis)

- Synthesize pairs of corresponding images
- Enforce weight-sharing constraints in high-level layers

---

UNIT (Unsupervised Cross-domain Image Translation)

UNIT learns translation functions of mapping an image in one domain to another without any corresponding images across domains.

Table 2: UDA results on adapting from the SVHN domain to the MNIST domain. The results of the other algorithms were duplicated from (Taigman et al., 2017)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA (Fernando et al., 2013)</td>
<td>59.32%</td>
</tr>
<tr>
<td>DANN (Ganin et al., 2016)</td>
<td>73.85%</td>
</tr>
<tr>
<td>DTN (Taigman et al., 2017)</td>
<td>84.88%</td>
</tr>
<tr>
<td>UNIT (proposed)</td>
<td>90.53%</td>
</tr>
</tbody>
</table>

Proposed Method - Cross-Domain Disentanglement (CDD)

Figure: Overview of our proposed method
Proposed Method - Cross-Domain Disentanglement (CDD)
Generative Adversarial Network (GAN)

✓ Synthesize realistic images

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D (\tilde{X})) + \mathbb{E}[\log D (X))]$$

Synthesized images $\tilde{X}$  
Real images $X$
Proposed Method
AuxiliaryClassifier-GAN (AC-GAN)

✓ _Synthesize_ images conditioned on disentangled factor $l$
✓ Disentangle the specific factor $l$ from the representation $z$

$$L_{dis} = \mathbb{E}[\log(L = l|X)] + \mathbb{E}[\log(L = l|\tilde{X})]$$
Proposed Method

VAE + AC-GAN

\[
\mathcal{L}_{VAE} = \mathcal{L}_{perc} + KL(q_S(z|X)||p(z))
\]
\[
\mathcal{L}_{perc} = \|\Phi(X) - \Phi(\tilde{X})\|_F^2
\]

✓ **Translate** the images conditioned on disentangled factor \( l \)
Proposed Method

VAE + AC-GAN

- Divide the network into low-level and high-level layers.
Proposed Method

VAE + AC-GAN

Divide the network into low-level and high-level layers.
Proposed Method

VAE + AC-GAN for cross-domain images

- Share the high-level layers of Encoder, Generator, and Discriminator

Source Domain

Target Domain

Joint Space
Proposed Method
VAE + AC-GAN for cross-domain images

- No attribute supervision in the target domain.
- We only urge the synthesized data in target domain $\tilde{X}_T$ to be disentangled.

$$\mathcal{L}_{dis}^T = \mathbb{E}[\log(L = l|\tilde{X}_T)]$$
Proposed Method

VAE + AC-GAN for cross-domain images

 ✓ Translate the images $\tilde{X}_{T \rightarrow S}$ and $\tilde{X}_{S \rightarrow T}$ across different domains.

\[
\mathcal{L}_{adv}^{cd} = \mathcal{L}_{adv}^{S \rightarrow T} + \mathcal{L}_{adv}^{T \rightarrow S},
\]
\[
\mathcal{L}_{adv}^{S \rightarrow T} = \mathbb{E}[\log(1 - D_C(D_T(\tilde{X}_{S \rightarrow T}))) + \mathbb{E}[\log(D_C(D_T(X_T)))],
\]
\[
\mathcal{L}_{adv}^{T \rightarrow S} = \mathbb{E}[\log(1 - D_C(D_S(\tilde{X}_{T \rightarrow S}))) + \mathbb{E}[\log(D_C(D_T(X_S)))].
\]
Proposed Method

VAE + AC-GAN for cross-domain images

![Diagram of VAE + AC-GAN for cross-domain images]

- **Source Domain**: $X_S$ to $E_S$, $E_C$ to $l$, $z$ to $G_C$, $G_S$ to $X_S$
- **Target Domain**: $X_T$ to $E_T$, $l$ to $X_T$, $D_T$, $D_S$ to $X_T$
- **Joint Space**: $l = 1, 2, ..., L$ for distinguishing between real and fake data

- **Encoder**: $E_S$, $E_C$
- **Generator**: $G_C$, $G_S$
- **Discriminator**: $D_S$, $D_T$

- **Objective**: Tie the disentangled factor $l$ across domains with

$$\mathcal{L}_{\text{dis}} = \mathbb{E}[\log(L = l|\tilde{X}_{S\rightarrow T})] + \mathbb{E}[\log(L = l|\tilde{X}_{T\rightarrow S})].$$
Proposed Method

VAE + AC-GAN for cross-domain images

Overall objective function can be defined as:

\[
\begin{align*}
\mathcal{L}_E &= \mathcal{L}_{VAE} \\
\mathcal{L}_G &= \mathcal{L}_{perc}^S + \mathcal{L}_{perc}^T + \mathcal{L}_{dis} + \mathcal{L}_{adv} \\
\mathcal{L}_D &= \mathcal{L}_{dis} - \mathcal{L}_{adv}
\end{align*}
\]
Proposed Method

VAE + AC-GAN for cross-domain images

✓ Overall objective function can be defined as:

\[ \mathcal{L}_E = \mathcal{L}_{VAE} \]
\[ \mathcal{L}_G = \mathcal{L}_S^{perc} + \mathcal{L}_T^{perc} + \mathcal{L}_{dis} + \mathcal{L}_{adv} \]
\[ \mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv} \]
Proposed Method

VAE + AC-GAN for cross-domain images

✓ Overall objective function can be defined as:

\[ \mathcal{L}_E = \mathcal{L}_{VAE} \]
\[ \mathcal{L}_G = \mathcal{L}_{perc}^S + \mathcal{L}_{perc}^T + \mathcal{L}_{dis} + \mathcal{L}_{adv} \]
\[ \mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv} \]
Proposed Method

VAE + AC-GAN for cross-domain images

Overall objective function can be defined as:

\[ \mathcal{L}_E = \mathcal{L}_{VAE} \]
\[ \mathcal{L}_G = \mathcal{L}_{perc}^S + \mathcal{L}_{perc}^T + \mathcal{L}_{dis} + \mathcal{L}_{adv} \]
\[ \mathcal{L}_D = \mathcal{L}_{dis} - \mathcal{L}_{adv} \]
Experiments

- Qualitative Evaluation:
  - Conditional image synthesis and translation

- Quantitative Evaluation:
  - Cross-domain attribute classification

- Dataset
  - CelebFaces Attributes dataset (CelebA)
  - A large-scale face dataset with 200K+ celebrity images with 40 facial annotated attributes
Results

S: faces w/o eyeglasses; T: faces w/ eyeglasses; l: attribute of smiling

Table 1: Cross-domain classification results of face images with respect to the attribute of smiling. Source and target-domain test data are faces w/o eyeglasses and w/ eyeglasses, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoGAN</td>
<td>87.03</td>
<td>71.92</td>
</tr>
<tr>
<td>UNIT</td>
<td>88.66</td>
<td>71.82</td>
</tr>
<tr>
<td>Ours*</td>
<td>89.48</td>
<td>88.43</td>
</tr>
<tr>
<td>Ours</td>
<td>89.73</td>
<td>84.43</td>
</tr>
</tbody>
</table>
Results

$S$ : real photo of faces; $T$ : simulated sketch of faces; $l$ : attribute of smiling

Table 2: Cross-domain classification results of face images with respect to the attribute of smiling. Source and target-domain test data are sketch and photo faces, respectively.
Results

S : real photo of faces; T : simulated sketch of faces; l : attribute of eyeglasses

Table 3: Cross-domain classification results of face images with respect to the attribute of eyeglasses. Source and target-domain test data are sketch and photo faces, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>CoGAN</th>
<th>UNIT</th>
<th>Ours*</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>96.63</td>
<td>97.65</td>
<td>97.06</td>
<td>97.19</td>
</tr>
<tr>
<td>Target</td>
<td>81.01</td>
<td>79.89</td>
<td><strong>94.49</strong></td>
<td><strong>94.84</strong></td>
</tr>
</tbody>
</table>
Summary

• Transfer Learning for
  • Homogeneous/heterogeneous domain adaptation
  • Multi-label classification / zero-shot learning
  • Robust face recognition (e.g., cross-resolution, cross-modality, etc.)

• Feature Disentanglement for
  • Cross-domain image synthesis/translation/classification
  • Only label supervision from a single (source) domain is needed