Adaptation for Objects and Attributes

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Adriana Kovashka (UT Austin),
Boqing Gong (USC), and Fei Sha (USC)
Learning-based visual recognition

Last 10+ years: impressive strides by learning appearance models (usually discriminative).

Annotator

Training images

New image

Image features
Typical assumptions

1. Test set will look like the training set.
2. Human labelers “see” the same thing.
Mismatched domains

TRAIN: Flickr

TEST: YouTube
Mismatched domains

TRAIN

Catalog images

TEST

Mobile phone photos
Mismatched domains

TRAIN

ImageNet

TEST

PASCAL VOC
“It is worthwhile to note that, even with 140K training ImageNet images, we do not perform as well as with 5K PASCAL VOC training images.”

– Perronnin et al. CVPR 2010
Mismatched domains

Problem: Poor cross-domain generalization
  • Different underlying distributions
  • Overfit to datasets’ idiosyncrasies

Possible solution:
  Unsupervised domain adaptation
Unsupervised domain adaptation

Setup

Source domain (with labeled data)

\[ D_S = \{(x_m, y_m)\}_{m=1}^{M} \sim P_S(X, Y) \]

Target domain (no labels for training)

\[ D_T = \{(x_n, y_n)\}_{n=1}^{N} \sim P_T(X, Y) \]

Objective

Learn classifier to work well on the target

Different distributions
Much recent research

Correcting *sampling* bias

- [Shimodaira, '00]
- [Huang et al., Bickel et al., '07]
- [Sugiyama et al., '08]
- [Sethy et al., '06]
- [Sethy et al., '09]

Adjusting mismatched *models*

+ [Evgeniou and Pontil, '05]
- [Duan et al., '09]
- [Duan et al., Daumé III et al., Saenko et al., '10]
- [Kulis et al., Chen et al., '11]

Inferring domain-invariant *features*

+ [Pan et al., '09]
- [Argyriou et al., '08]
[Daumé III, '07]
[Blitzer et al., '06]
+ [Gopalan et al., '11]
[Muandet et al., '13]
-Gong et al., '12
[Chen et al., '12]

- [This work]
Problem

Existing methods attempt to adapt all source data points, including “hard” ones.
Problem

Existing methods attempt to adapt all source data points, including “hard” ones.

Our idea

Automatically identify the “most adaptable” instances

Use them to create series of easier auxiliary domain adaptation tasks

[Gong et al., ICML 2013]
Landmarks are labeled source instances distributed similarly to the target domain.

[Gong et al., ICML 2013]
Landmarks are labeled source instances distributed similarly to the target domain.

Roles:

- Ease adaptation difficulty
- Provide discrimination (biased to target)

[Gong et al., ICML 2013]
1. Identify landmarks

at multiple scales.

[Gong et al., ICML 2013]
Key steps

1. Construct auxiliary domain adaptation tasks

2. Obtain domain-invariant features

3. Predict target labels

[Gong et al., ICML 2013]
Identifying landmarks

Objective

$$P_L(\text{landmarks}) \approx P_T(\text{target})$$

$$\min_{\text{landmarks}} d(P_L, P_T) ?$$

[Gong et al., ICML 2013]
Maximum mean discrepancy (MMD)

**Empirical estimate** [Gretton et al. ’06]

\[
d(P_L, P_T) = \left\| \frac{1}{L} \sum_{l=1}^{L} \phi(x_l) - \frac{1}{N} \sum_{n=1}^{N} \phi(x_n) \right\|_\mathcal{H}
\]

\(\mathcal{H}\) a universal RKHS

\(\phi(\cdot)\) kernel function induced by \(\mathcal{H}\)

\(x_l\) the \(l\)-th landmark (from the source domain)

[Gong et al., ICML 2013]
Method for identifying landmarks

Integer programming

$$\min_{\{\alpha_m\}} \left\| \frac{1}{\sum_i \alpha_i} \sum_{m=1}^{M} \alpha_m \phi(x_m) - \frac{1}{N} \sum_{n=1}^{N} \phi(x_n) \right\|^2_{\mathcal{H}}$$

where

$$\alpha_m = \begin{cases} 1 & \text{if } x_m \text{ is a landmark for the target} \\ 0 & \text{else} \end{cases}$$

$$m = 1, 2, \cdots, M$$

[Gong et al., ICML 2013]
Method for identifying landmarks

Convex relaxation

\[
\min_{\{\alpha_m\}} \left\| \frac{1}{\sum_i \alpha_i} \sum_{m=1}^{M} \alpha_m \phi(x_m) - \frac{1}{N} \sum_{n=1}^{N} \phi(x_n) \right\|_{\mathcal{H}}^2
\]

\[
\beta_m = \frac{\alpha_m}{\sum_i \alpha_i} \rightarrow \text{Quadratic programming}
\]

\[
\min_{\beta} \beta^T K^s \beta - \frac{2}{N} \beta^T K^{st} 1
\]

[Gong et al., ICML 2013]
Scale for landmark similarity?

$$\min_{\beta} \beta^T K^s \beta - \frac{2}{N} \beta^T K^{st} 1$$

Gaussian kernels

How to choose the bandwidth?

**Our solution:**

Examine distributions at **multiple** granularities

Multiple bandwidths $\rightarrow$ multiple sets of landmarks

[Gong et al., ICML 2013]
Landmarks at multiple scales

<table>
<thead>
<tr>
<th>Target</th>
<th>Headphone</th>
<th>Mug</th>
</tr>
</thead>
<tbody>
<tr>
<td>target</td>
<td><img src="image1" alt="Headphones" /></td>
<td><img src="image2" alt="Mugs" /></td>
</tr>
<tr>
<td>$\sigma=2^6$</td>
<td><img src="image3" alt="Headphones" /></td>
<td><img src="image4" alt="Mugs" /></td>
</tr>
<tr>
<td>$\sigma=2^0$</td>
<td><img src="image5" alt="Headphones" /></td>
<td><img src="image6" alt="Mugs" /></td>
</tr>
<tr>
<td>$\sigma=2^{-3}$</td>
<td><img src="image7" alt="Headphones" /></td>
<td><img src="image8" alt="Mugs" /></td>
</tr>
<tr>
<td>Unselected</td>
<td><img src="image9" alt="Headphones" /></td>
<td><img src="image10" alt="Mugs" /></td>
</tr>
</tbody>
</table>

[Gong et al., ICML 2013]
Key steps

2 Construct auxiliary domain adaptation tasks

Constructing easier auxiliary tasks

At each scale $\sigma$

New source $= \text{Source} \setminus \text{Landmarks}$

New target $= \text{Target} \cup \text{Landmarks}$

Intuition: distributions are closer (cf. Theorem 1)

[Gong et al., ICML 2013]
At each scale $\sigma$

New source = Source \ Landmarks

New target = Target \cup Landmarks

Intuition: distributions are closer (cf. Theorem 1)

[Gong et al., ICML 2013]
Constructing easier auxiliary tasks

Each task provides new basis of features via geodesic flow kernel (GFK):

\[ K_\sigma(x_i, x_j) = \int_0^1 (\Phi_\sigma(t)'x_i)'(\Phi_\sigma(t)'x_j)dt = x_i G_\sigma x_j \]

- Integrate out domain changes

[Gong et al., CVPR 2012]
Key steps

2 Construct auxiliary domain adaptation tasks

3 Obtain domain-invariant features

\[ \Phi(x) = \begin{bmatrix} \Phi_1(x) \cdot w_1 \\ \Phi_2(x) \cdot w_2 \\ \Phi_3(x) \cdot w_3 \end{bmatrix} \]

MKL
Combining features discriminatively

Multiple kernel learning on the labeled landmarks

\[ F = \sum_{\sigma} w_\sigma G_\sigma, \quad \text{s.t.} \quad w_\sigma \geq 0, \sum_{\sigma} w_\sigma = 1 \]

Arriving at domain-invariant feature space

Discriminative loss biased to the target
Key steps

2. Construct auxiliary domain adaptation tasks

3. Obtain domain-invariant features

4. Predict target labels

\[ \Phi(x) = \begin{bmatrix} \Phi_1(x) \cdot w_1 \\ \Phi_2(x) \cdot w_2 \\ \Phi_3(x) \cdot w_3 \end{bmatrix} \]
Experiments

Four vision datasets/domains on visual object recognition
[Griffin et al. ’07, Saenko et al. 10’]

Four types of product reviews on sentiment analysis
Books, DVD, electronics, kitchen appliances [Biltzer et al. ’07]
Cross-dataset object recognition

Accuracy (%)

No adaptation

A→C   A→D   C→A   C→W   W→A   W→C
Cross-dataset object recognition

![Accuracy chart showing performance comparison between methods and datasets.](image)
Cross-dataset object recognition

Accuracy (%)
Datasets as domains?

Domain 1

Domain 2

ASSUMED

Domain 3

Domain 4

Domain 5
Datasets as domains?
Datasets as domains?

Dataset != Domain

Cross-dataset adaptation is suboptimal
How to define a domain?

**NLP:** *Language*-specific domains ✓

**Speech:** *Speaker*-specific domains ✓

**Vision:** ??

- pose-specific?
- *illumination*-specific?
- occlusion?
- *image resolution*?
- background?

**Challenges:**

- Many continuous factors vs. few discrete factors
- Factors overlap and interact
Discovering latent visual domains

We propose to **discover** domains – “reshaping” them to cross dataset boundaries

**Maximum distinctiveness**

\[
\max_{\{z_{m,k}\}} \sum_{k \neq k'} d(P(k), P(k')) \rightarrow \text{MMD}
\]

where \( z_{m,k} = \begin{cases} 
1 & \text{if } x_m \text{ belongs to domain } k \\
0 & \text{else}
\end{cases} \)

**Maximum learnability**

Determine \( K \) with domain-wise cross-validation

[Gong et al., NIPS 2013]
Results: discovering domains

Discovered domain I

Discovered domain II

[Gong et al., NIPS 2013]
Results: discovering domains

Cross-dataset object recognition

Accuracy

<table>
<thead>
<tr>
<th>Domains= datasets</th>
<th>Hoffman et al. 2012</th>
<th>Discovered domains (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

Cross-viewpoint action recognition

Accuracy

<table>
<thead>
<tr>
<th>Domains= datasets</th>
<th>Hoffman et al. 2012</th>
<th>Discovered domains (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>
Summary so far

**landmarks**

- labeled source instances distributed similarly to the target
- auxiliary tasks provably easier to solve
discriminative loss despite unlabeled target

**reshaping datasets to latent domains**
discover cross-dataset domains
maximally distinct & learnable
Typical assumptions

1. Test set will look like the training set.
2. Human labelers “see” the same thing.
Visual attributes

• High-level semantic properties shared by objects
• Human-understandable and machine-detectable

Standard approach

Learn one monolithic model per attribute
Problem

There may be valid perceptual differences within an attribute.

Formal? User labels: 50% “yes” 50% “no”

More ornamented? User labels: 50% “first” 20% “second” 30% “equally”

Binary attribute

Relative attribute
Imprecision of attributes

Fine-grained meaning

Overweight?
or just
Chubby?
Imprecision of attributes

Context

Is \( \textit{formal} \)?

= \textit{formal} wear for a \textit{conference}? OR

= \textit{formal} wear for a \textit{wedding}?
Imprecision of attributes

Cultural

Is blue or green?

English: “blue”

Russian: “neither”
(“голубой” vs. “синий”)

Japanese: “both”
(“青” = blue and green)
But do we need to be that precise?

Yes. Applications like image search require that user’s perception matches system’s predictions.

“white high heels”

“less formal than these”

[WhittleSearch, Kovashka et al. CVPR 2012]
Our idea

• Treat learning perceived attributes as an adaptation problem.
• Adapt generic attribute model with minimal user-specific labeled examples.
• Obtain implicit user-specific labels from user’s search history

[Kovashka and Grauman, ICCV 2013]
Our idea

Vote on labels

"formal"

"not formal"

Adapt

"formal"

"not formal"

[Kovashka and Grauman, ICCV 2013]
Learning adapted attributes

• Adapting binary attribute classifiers:

Given user-labeled data \( D_b = \{ x_i, y_i \}_{i=1}^{N} \) and generic model \( \omega'_b \),

\[
\min_{\omega_b} \frac{1}{2} \| \omega_b - \omega'_b \|^2 + C \sum_{i=1}^{N} \xi_i,
\]

subject to \( y_i x_i^T \omega_b \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \)

J. Yang et al. ICDM 2007.
Learning adapted attributes

• Adapting relative attribute rankers:

Given user-labeled data

\[ D_r = \{ (x_{i_1} \succ x_{j_1}) \}_{i=1}^{N} \]

and generic model \( w_r' \),

\[
\min_{w_r} \frac{1 - \delta}{2} \| w_r \|^2 + \frac{\delta}{2} \| w_r - w_r' \|^2 + C \sum_{i=1}^{N} \xi_i
\]

subject to

\[ w_r^T x_{i_1} - w_r^T x_{i_2} \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i, \]

Collecting user-specific labels

- **Explicitly** from actively requested labels
  Seek labels on uncertain and diverse images

- **Implicitly** from search history
  - Transitivity
    “My target is…
    less formal than
    more formal than
    “
  - Contradictions
Inferring implicit labels

“Target is more sporty than B”

“Target is less sporty than A”

User’s feedback history can reveal mismatch in perceived and predicted attributes
Inferring implicit labels

“Target is more sporty than B”

more feminine (~ less sporty) more sporty

“Target is more feminine than A”

User’s feedback history can reveal mismatch in perceived and predicted attributes
Datasets

SUN Attributes:
- 14,340 scene images
- 12 attributes: “sailing”, “hiking”, “vacationing”, “open area”, “vegetation”, etc.

Shoes:
- 14,658 shoe images;
- 10 attributes: “pointy”, “bright”, “high-heeled”, “feminine” etc.
Adapted attribute accuracy

- 3 datasets
- 22 attributes
- 75 total users
Adapted attribute accuracy

- 3 datasets
- 22 attributes
- 75 total users
Adapted attribute accuracy

- 3 datasets
- 22 attributes
- 75 total users
Adaptation approach most accurately captures perceived attributes

[Kovashka and Grauman, ICCV 2013]
Which images most influence adaptation?

- pointy
- open
- bright
- ornamented
- shiny
- high-heeled
- long
- formal
- sporty
- feminine
- sailing
- vacationing
- hiking
- camping
- socializing
- shopping
- vegetation
- clouds
- natural light
- cold
- open area
- horizon far
Visualizing adapted attributes

Shoes – Relative Attributes – “Formal”

SUN – Binary Attributes – “Vacationing”
Personalizing image search with adapted attributes

“white shiny heels”

“shinier than”

Match rate

0 10 20 30 40 50 60 70

Shoes-Binary

SUN

- generic
- generic+
- user-exclusive
- user-adaptive
Impact of implicit labels

“Target is more sporty than B”
more feminine (≈ less sporty) more sporty

“Target is more feminine than A”

Percentile rank

Shoes-Relative
- explicit labels only
- +contradictions
- +transitivity
Summary

• Practical concerns if learning visual categories:
  Test images can look different from training images!
  People do not perceive image labels universally!

• Domain adaptation methods help address them
  Landmark-based unsupervised adaptation
  Reshaping datasets into latent domains
  Adapt generic models to account for user-specific perception of attributes
References


