1 Motivation

2 Historical points

3 Definition

4 Case studies
Outline

1. Motivation
2. Historical points
3. Definition
4. Case studies
Standard Supervised Learning Task

Most ML tasks assume the training/test data are drawn from the same data space and the same distribution.
Most ML tasks assume the training/test data are drawn from the same data space and the same distribution.
NLP tasks: POS, NER, Category labelling

Modified from Gao et al.'s presentation in KDD '08
Combine and get better result

Modified from Gao et al.’s presentation in KDD ’08
Traditional ML tasks assume the training/test data are drawn from the same data space and the same distribution.

Insufficient labelled data result in poor prediction performance.
- Lots of (un-)related existing data from various sources.

Start from scratch is always time-consuming.

Transfer knowledge from other sources may help!
Motivation (Taylor et. al JMLR ’09)
Outline

1 Motivation
2 Historical points
3 Definition
4 Case studies
In 1901, Thorndike and Woodworth explored how individuals transfer similar characteristics shared by different contexts.
In 1901, Thorndike and Woodworth explored how individuals transfer similar characteristics shared by different contexts.

In 1992, Perkins and Salomon published ”Transfer of Learning” which defined different types of transfer.
In 1901, Thorndike and Woodworth explored how individuals transfer similar characteristics shared by different contexts.

In 1992, Perkins and Salomon published “Transfer of Learning” which defined different types of transfer. Examples:

- **Skill learning:** $C/C++ \rightarrow Python$
- **Language acquisition:** $German \rightarrow English$
- Explanation-Based Neural Network Learning: A **Lifelong Learning** Approach [Thrun PhD ’95, NIPS ’96]
Machine Learning

- **Explanation-Based Neural Network Learning**: A *Lifelong Learning* Approach [Thrun PhD ’95, NIPS ’96]
- **Multitask Learning** [Caruana ICML ’93 & ’96, PhD ’97]
Machine Learning

- Explanation-Based Neural Network Learning: A *Lifelong Learning* Approach [Thrun PhD ’95, NIPS ’96]
- *Multitask Learning* [Caruana ICML ’93 & ’96, PhD ’97]
- Workshops
  - Learning to Learn: Knowledge Consolidation and Transfer in Inductive Systems [NIPS ’95]
  - Inductive Transfer: 10 Years Later [NIPS ’05]
  - Structural Knowledge Transfer for Machine Learning [ICML ’06]
  - Transfer Learning for Complex Tasks [AAAI ’08]
  - Lifelong Learning [AAAI ’11]
  - Theoretically Grounded Transfer Learning [ICML ’13]
  - Workshop: Second Workshop on Transfer and Multi-Task Learning: Theory meets Practice [NIPS ’14]
  - ...
Outline

1. Motivation
2. Historical points
3. Definition
4. Case studies
Definition

**Notations**

- **Domain $\mathcal{D}$**
  1. Data space $\mathcal{X}$
  2. Marginal distribution $P(X)$, where $X \in \mathcal{X}$

- **Task $\mathcal{T}$ (Given $\mathcal{D} = \{\mathcal{X}, P(X)\}$)**
  1. Label space $\mathcal{Y}$
  2. Learn a $f : X \rightarrow Y$ to approach the underlying $P(Y|X)$, where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
Assume we have only one source $S$ and one target $T$:

**Definition**

Transfer Learning (TL): Given a source domain $D_S$ and learning task $T_S$, a target domain $D_T$ and learning task $T_T$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $D_T$ using the knowledge in $D_S$ and $T_S$, where

$$D_S \neq D_T \quad (\text{either } X_S \neq X_T \text{ or } P_S(X) \neq P_T(X))$$

or

$$T_S \neq T_T \quad (\text{either } Y_S \neq Y_T \text{ or } P(Y_S|X_S) \neq P(Y_T|X_T))$$
Example: Category labelling

The Wall Street Journal
Chinese Version

Predictive Model

Science News

Pei-Hao (Eddy) Su and Yingzhen Li
Transfer Learning
Example: Category labelling

The Wall Street Journal
Chinese Version

Data $X_S$: $X_S \in \mathbb{R}^N$
$N$: Chinese lexicon size

$P_S(X) \neq P_T(X)$

ScienceNews

Data $X_T$: $X_T \in \mathbb{R}^M$
$M$: English lexicon size

Pei-Hao (Eddy) Su and Yingzhen Li
Transfer Learning
Example: Category labelling

The Wall Street Journal
Chinese Version

ScienceNews

Labels $Y_S$:
1. Markets
2. Economy
3. Management
4. Politics
$|Y_S| = 4$

Predictive Model

$P(Y_S|X_S) \neq P(Y_T|X_T)$

Labels $Y_T$:
1. Math & Tech
2. Body & Brain
3. Energy
4. Cosmology
5. Genes
$|Y_T| = 5$
ML v.s. TL (Langley ’06, Yang et al. ’13)
Outline

1. Motivation
2. Historical points
3. Definition
4. Case studies
The rest of the talk will give you an intuition, with examples, on:

- **when** to transfer
- **what** to transfer
- **and how** to transfer
When to transfer: Domain relatedness

Transfer learning is applicable when there exists relatedness

- Standard machine learning assume source = target
- Transferring knowledge from unrelated domain can be harmful
  - Negative transfer [Rosenstein et al NIPS-05 Workshop]
- (Ben-David et al.) proposed a bound of target domain error

Reference

Ben-David et al. Analysis of Representation for Domain Adaptation. NIPS '06
When to transfer (Ben-David et al.)

In standard binary classification supervised learning task:
- Given $X, Y = \{0, 1\}$ and samples from $P(x, y)$, we aim to learn $f : X \rightarrow [0, 1]$ which captures $P(y|x)$
- Often we decompose the problem into:
  1. determine a feature mapping $\Phi : X \rightarrow Z$
  2. learn a hypothesis $h : Z \rightarrow \{0, 1\}$ on dataset $\{\Phi(x), y\}$

In transfer learning scenario:

**Theorem (Simplified version of Thm. 1&2)**

Given $X = X_S = X_T$ and $P_S(x), P_T(x)$ the distributions of the source and target domain. Let $\Phi : X \rightarrow Z$ be a fixed mapping function and $\mathcal{H}$ be a hypothesis space. For any hypothesis $h \in \mathcal{H}$ trained on source domain:

$$\epsilon_T(h) \leq \epsilon_S(h) + d_\mathcal{H}(\tilde{P}_S, \tilde{P}_T) + \epsilon_S(h^*) + \epsilon_T(h^*)$$

where $\tilde{P}_S, \tilde{P}_T$ are induced distributions on $Z$ wrt. $P_S$ and $P_T$,

$h^* = \arg \min_{h \in \mathcal{H}} (\epsilon_S(h) + \epsilon_T(h))$ is the best hypothesis by joint training.
Domain adaptation

Approach 1: mixture of general & specific component

Can we learn hypotheses for both the general and specific components?

Reference:
Daume III. Frustratingly easy domain adaptation. ACL ’07
Daume III et al. Co-regularization Based Semi-supervised Domain Adaptation. NIPS ’10
EasyAdapt (Daume III)

Binary classification problem:
- \( X_S = X_T \subset \mathcal{R}^d, \ Y_S = Y_T = \{-1, +1\} \)
- Goal: obtain classifier \( f_T: X_T \rightarrow Y_T \)
- in SVM context: learn a hypothesis \( h_T \in \mathcal{R}^d \)

However:
- too little training data available on \((X_T, Y_T)\) for robust training
- also \(P(x_S) \neq P(x_T)\) and \(P(x_S, y_T) \neq P(x_S, y_T)\)
- ...so directly apply a trained hypothesis \( h_S \) returns bad results

How to use \( x_S, y_S \sim P(x_S, y_S) \) to improve learning of \( h_T \)?
EasyAdapt algorithm

- define two mappings $\Phi_S, \Phi_T : \mathcal{R}^d \to \mathcal{R}^{3d}$:
  
  $\Phi_S(x_S) = (x_S, x_S, 0), \quad \Phi_T(x_T) = (x_T, 0, x_T)$

- training: learn a hypothesis $h = (w^g, w^s, w^t) \in \mathcal{R}^{3d}$ on transformed dataset $\{(\Phi_S(x_S), y_S)\} \cup \{(\Phi_T(x_T), y_T)\}$

- test: apply $h_T = w^g + w^t$ on $x_T$

- (also $h_S = w^g + w^s$)
Use unlabelled data to improve training:

- want $h_S$ and $h_T$ to agree on unlabelled data $x_U$:

$$h_S \cdot x_U = h_T \cdot x_U \iff w^s \cdot x_U = w^t \cdot x_U \iff h \cdot (0, x_U, -x_U) = 0$$

- so we define mapping $\Phi_U : \mathcal{R}^d \to \mathcal{R}^{3d}$ for unlabelled data

$$\Phi_U(x_U) = (0, x_U, -x_U) \quad (1)$$

- and train the hypothesis $h$ on augmented and transformed dataset

$$\{(\Phi_S(x_S), y_S)\} \cup \{(\Phi_T(x_T), y_T)\} \cup \{\Phi_U(x_U), 0)\$$
**Outline**

Motivation

Historical points

Definition

Case studies

---

**EA++ (Daume III et al.)**

(a) DVD → BOOKS (proxy A-distance=0.7616),
(b) KITCHEN → APPAREL (proxy A-distance=0.0459).

---

- **SOURCE/TARGETONLY(-FULL):** trained on source/target (full) labelled samples
- **ALL:** trained on combined labelled samples
- **EA/EA++:** trained in augmented feature space (and unlabelled target data)
Feature transfer

Approach 2: shared lower-level features

- DNN first layer learns Gabor filters or color blobs when trained on images
- instances in source/target domain share the same lower-level features?

Reference:
Feature transfer

Lee et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. ICML ’09

adapt from Ruslan Salakhutdinov’s tutorial in MLSS’14 Beijing
Feature transfer (Yosinski et al.)

Pei-Hao (Eddy) Su and Yingzhen Li
Feature transfer (Yosinski et al.)

Test 1 (similar datasets): random A/B splits of the ImageNet dataset

(similar source and target domain training/testing instances)
Feature transfer (Yosinski et al.)

Test 1 (similar datasets): random A/B splits of the ImageNet dataset

(similar source and target domain training/testing instances)
Feature transfer (Yosinski et al.)

Test 1 (similar datasets): random A/B splits of the ImageNet dataset

(similar source and target domain training/testing instances)
Feature transfer (Yosinski et al.)

Test 2 (very different datasets): man-made/natural object split

(dissimilar source and target domain training/testing instances)
Feature transfer (Yosinski et al.)

Test 2 (very different datasets): man-made/natural object split

(dissimilar source and target domain training/testing instances)
Joint representation

Approach 3: joint feature representation

- data has many domain specific characteristics
- however might be related in high level?
- our brain might work like this as well

Reference:
Joint representation (Srivastava et al.)

MIR Flickr Dataset http://press.liacs.nl/mirflickr/

<table>
<thead>
<tr>
<th>Classes</th>
<th>Images</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>baby, female, people, portrait</td>
<td><img src="image1.png" alt="Image" /></td>
<td>claudia</td>
</tr>
<tr>
<td>plant life, river, water</td>
<td><img src="image2.png" alt="Image" /></td>
<td>(no text)</td>
</tr>
<tr>
<td>clouds, sea, sky, transport, water</td>
<td><img src="image3.png" alt="Image" /></td>
<td>barco, pesca, boattosail, navegacao</td>
</tr>
<tr>
<td>animals, dog, food</td>
<td><img src="image4.png" alt="Image" /></td>
<td>watermelon, hilarious, chihuahua, dog</td>
</tr>
<tr>
<td>clouds, sky, structures</td>
<td><img src="image5.png" alt="Image" /></td>
<td>colors, cores, centro, comercial, building</td>
</tr>
</tbody>
</table>

- For images
  - 1M datapoints, 25K labelled instances in 38 classes, 10K for training, 5K for validation and 10K for testing
  - inputs are the concatenation of PHOW and MPEG-7 features
- For texts
  - use word count vectors on 2K frequently used tags (very sparse)
  - 18% training images have missing texts
for images: 2-layer deep Boltzmann machine (DBM) with Gaussian input units ($v_{mi} \in \mathbb{R}$, abbrev. $W^{(k)}_m(i,j)$ as $W_{ij}^{(k)}$)

$$P(v_m, h^{(1)}_m, h^{(2)}_m) \propto \exp \left( - \sum_i \frac{(v_{mi} - b_i)^2}{2\sigma_i^2} + \sum_{i,j} \frac{v_{mi}}{\sigma_i} W_{ij}^{(1)} h^{(1)}_{mj} + \sum_{j,l} h^{(1)}_{mj} W_{jl}^{(2)} h^{(2)}_{ml} \right)$$
Joint representation (Srivastava et al.)

for texts: 2-layer DBM with replicated softmax model ($v_{ti}$ counts the occurrence of word $i$, abbrev. $W_t^{(k)}(i,j)$ as $W_{ij}^{(k)}$)

$$P(v_t, h^{(1)}_t, h^{(2)}_t) \propto$$

$$\exp \left(- \sum_{i=1} v_{ti} b_i + \sum_{i,j} v_{ti} W_{ij}^{(1)} h_{mj}^{(1)} + \sum_{j,l} h_{ij}^{(1)} W_{jl}^{(2)} h_{tl}^{(2)} \right)$$
Joint representation (Srivastava et al.)

Combining domain specific models to a multimodal DBM:

\[ P(v_m, v_t, h; \theta) \propto \exp \left( -E(h_m^{(2)}, h_t^{(2)}, h^{(3)}) - E(v_m, h_m^{(1)}, h_m^{(2)}) - E(v_t, h_t^{(1)}, h_t^{(2)}) \right) \]
Joint representation (Srivastava et al.)

- first pre-train domain specific DBMs with CD, then co-train the joint model with PCD
- use mean-field variational approximation when computing hidden unit moments driven by data
Joint representation (Srivastava et al.)

Results:

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP</th>
<th>Prec@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.124</td>
<td>0.124</td>
</tr>
<tr>
<td>SVM (Huiskes et al., 2010)</td>
<td>0.475</td>
<td>0.758</td>
</tr>
<tr>
<td>LDA (Huiskes et al., 2010)</td>
<td>0.492</td>
<td>0.754</td>
</tr>
<tr>
<td>DBM</td>
<td>0.526 ± 0.007</td>
<td>0.791 ± 0.008</td>
</tr>
<tr>
<td>DBM (using unlabelled data)</td>
<td>0.585 ± 0.004</td>
<td>0.836 ± 0.004</td>
</tr>
</tbody>
</table>

**Figure:** Classification with data from both image and text domain

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP</th>
<th>Prec@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image LDA (Huiskes et al., 2010)</td>
<td>0.315</td>
<td>-</td>
</tr>
<tr>
<td>Image SVM (Huiskes et al., 2010)</td>
<td>0.375</td>
<td>-</td>
</tr>
<tr>
<td>Image DBN</td>
<td>0.463 ± 0.004</td>
<td>0.801 ± 0.005</td>
</tr>
<tr>
<td>Image DBM</td>
<td>0.469 ± 0.005</td>
<td>0.803 ± 0.005</td>
</tr>
<tr>
<td>Multimodal DBM (generated text)</td>
<td>0.531 ± 0.005</td>
<td>0.832 ± 0.004</td>
</tr>
</tbody>
</table>

**Figure:** Classification with data from image domain only
Joint representation (Srivastava et al.)

Results:

Figure: Retrieval results for multi/image domain queries
In this talk, we showed that

- transfer learning adapts knowledge from other sources to improve target task performance
- domains related to each other in different ways

In the future:

- manage large scale data that do not lack in size but may lack in quality
- manage data which may continuously change over time
Open Questions

- what are the limits of existing multi-task learning methods when the number of tasks grows while each task is described by only a small bunch of samples ("big T, small n")?
- what is the right way to leverage over noisy data gathered from the Internet as reference for a new task?
- how can an automatic system process a continuous stream of information in time and progressively adapt for life-long learning?
- can deep learning help to learn the right representation (e.g., task similarity matrix) in kernel-based transfer and multi-task learning?
- How can similarities across languages help us adapt to different domains in natural language processing tasks?
- ...

\(^2\text{nips.cc/Conferences/2014/Program/event.php?ID=4282}\)
Thank you
References

9. Daume III. Frustratingly easy domain adaptation. ACL 2007