Domain Adaptation–
20 years of theory chasing practice
Subjective view

Shai Ben-David
University of Waterloo,
and Vector Institute, Toronto, Canada

Simons Domain Adaptation workshop
November, 2024

Transfer Learning

Common phenomenon:
The data generation at application time differs from the training data generation.

Common goal of transfer learning:
Adapt an existing model rather than retrain on the target task from scratch

Examples of transfer learning

- Domain Adaptation
  - Train predictor on one task; apply (or adapt to) a different task
  - Typical setting: lots of annotated data from a source task; mostly unlabeled data from a target task
- Single/Few shot learning
  - Domain adaptation with very few data points from the target task
- Multitask learning
  - Train a joint predictor/model for multiple tasks (cats, dogs, zebras..)
- Lifelong learning
  - Adapt a predictor continuously over time
- Adversarial learning
  - Often not considered a transfer task; we want the predictor to be perturbation robust

In this talk I will focus on Domain Adaptation
and binary classification.

My first encounter - Snowbird 2005

John Blitzer and Fernando Pereira approached me about their paper *"Domain adaptation with structural correspondence learning"*

They had a successful algorithm adapting a Part of Speech (POS) tagger trained on Wall Street Journal text to work on MEDLINE text.

They wanted to come up with a theory explaining its success.

Their POS Domain Adaptation paradigm

- Choose a set of “pivot words” (determiners, prepositions, connectors and frequently occurring verbs).
- Represent every word in a text as a vector of its correlations, within a small ‘window”, with each of the pivot words.
- Train a linear separator on the (images of) the training data coming from one domain and use it for tagging on the other.

Our abstraction [BD- Blitzer- Kramer-Pereira 2006]

Our proposed algorithmic paradigm:

1. Embed the original attribute space (of both the source and target distributions) into some feature space in which
   1.1 The two tasks look similar.
   1.2 The source task can still be well classified.
2. To predict: Represent your test point in that feature space, and use a source trained classifier to predict its label.

Our theoretical guarantee

Given a hypothesis class H, for every $h \in H$

$$L_Q(h) \leq L_P(h) + d_{H\Delta H}(P_X, Q_X) + \lambda$$

Where $d_{H\Delta H}(P_X, Q_X)$ is a measure of the discrepancy between the marginal distributions

and λ is a measure of the labeling disagreement.

Theory vs Practice view

This felt like a good match between theory and practice.

- The theoretical contribution -
  1. Explain what is a good representation.
  2. Once you have that, you do not need any adaptation.

- The practice paradigm there - Use prior domain knowledge to hand-craft a good representation.

Followup theory progress

Theory followed up on roughly three themes:

1. Generalize the measures of P - Q discrepancy.

2. Incorporating more target training data information:
   - Randomly generated target labeled samples.
   - Active querying of target labels.

3. Algorithms for finding good data representations.

Measures of Source-target discrepancy/relatedness

Starting with the work of **Yishay Mansour, Mehryar Mohri and Afshin Rostamizadeh**

Generalizing to various learning tasks (Multi-class, regression, novelty detection) and various losses.

Most recently the work of **Steve Hanneke and Samory Kpotufe**.

Main questions addressed:

1. What type of relatedness may help DA.
2. Performance guarantees as a function of relatedness.
3. Quantitative evaluation of relatedness.

Main open challenge - detection and measuring source-target discrepancies

The challenge of Change Detection

There are obvious No Free Lunch theorems, implying that there is no reliable way to detect distribution change.

In particular, no way to distinguish In-Distribution from Out-of-Distribution data.

(And in a way to distinguish Interpolation from Extrapolation generalization).

Incorporating target empirical data

Starting with **[BD, Blitzer, et al 2010]**:
Minimize Weighted Empirical Loss, where the weighing depends on the source-target discrepancy and the relative sample sizes.

Most recently **[Hanneke,Kpotufe, 2024]**:
Notions of Confidence Sets.

The balancing of source and target data heavily dependence in quantifying their relative discrepancy

The high level issues

Optimizing the incorporation of the source and target data can be viewed under the **fundamental issues** of

- Balancing the utilization prior knowledge and posterior evidence.
- **Confidence/uncertainty evaluation**.

The search for useful representations

This is where the theory begins to fade. There is much academic research between theory and practice:

Approaches to developing data representations for DA include:

- Contrastive Learning
- Causality and feature invariance:
  - The language of **causality** is somewhat misleading - gap between intuitive semantics and the technical use
  - **Invariance** is a clearer notion. However, there is no clear distinction between Invariant and spurious features.
- ..
- ...
- Foundational models

Academic research towards practical developments

Attempt to divide into three rough categories:

1. *N*atural sciences approach:
   - "Anthropological studies" - probing LLM's (**Jacob Steinhardt's** talk).
   - Design experiments on crafted data (e.g., manipulation of images and text) (**Tengyu Ma**
2. Developing practical heuristics (**Rich Zemel's** Few shot learning)
3. Tools that are actually used in real applications (**Zack Lipton**?).

Concluding messages

I think that there is no hope for a general task-independent theory of DA,

Impactful DA tools and analysis should focus on particular tasks:

- Images
- NLP
- Medical diagnosis
- Commercial applications (marketing, advertizing)
- Agricultural/environmental applications.
- Adapting to temporal change (aging?).
- Etc.