A Representation Learning Approach for Domain Adaptation

Pascal Germain

INRIA Lille

MODAL SEMINAR

November 7, 2016

*Every classification throws light on something.*

— Isaiah Berlin
Disclaimer

This talk does not contain any trace of PAC-Bayesian theory...
Joint Work with...

- François Laviolette
- Mario Marchand
- Hana Ajakan
- Hugo Larochelle
- Yaroslav Ganin
- Evgeniya Ustinova
- Victor Lempitsky

Université Laval, Québec, Canada

Université de Sherbrooke, Québec, Canada and Google Brain

Skolkovo Institute of Science and Technology, Moscow Region, Russia
Plan

1. Domain Adaptation Setting
2. Theoretical Foundations
3. Domain-Adversarial Neural Network (DANN)
4. Empirical Results with “Shallow” Networks
5. Empirical Results with “Deep” Networks
6. Derivative Works and New Ideas
7. Conclusion
Example

Book critics
(target)

- The end of the series.
  This book was written to provoke those who wanted Adams to continue the trilogy but I loved it. Author settled down on a bob fearing planet where he has acquired the prestigious...
  [Read more]
  Published on Mar 18 2002 by Dan

- Mostly Harmless is underrated
  I think most of the reviews for this book downplay it seriously. While the ending is kind of disappointing, the book overall is wonderful.
  [Read more]
  Published on Jan 22 2002 by A Big Adams Fan

- Please pretend this book was never written.
  I have long been a fan of the Hitchhikers series as they are comic genius. The book Mostly Harmless, however, should never have come about. It is frustration at its peak.
  [Read more]
  Published on Jan 14 2002 by Paul Norrod

- Kinda like horror movies...
  ...in that the last one usually isn't all that appealing. I liked it fine, with some of Adams's wit, but it was a bit disappointing.
  [Read more]
  Published on Nov 4 2001 by Kristopher Vincent

- A Terrible End to a Great Series
  The ending for this books was so bad that I vowed never to read another Douglas Adams book. Adams was obviously sick and tired of the series and used this book to kill it off with...
  [Read more]
  Published on Oct 17 2001 by David A. Lessnau

Movie critics
(source)

- Don't Panic!
  If you haven't listened to the BBC radio-play, this isn't bad! Purists, no doubt, will dispute my verdict but the fact of the matter is THGTTG (see title) does have Douglas Adams'...
  [Read more]
  Published on Mar 13 2011 by Sid Matheson

- On Blu-ray, even better
  I've seen this movie on TV and wanted to add it to my collection. I couldn't find it locally so when I saw it on amazon and on Blu-ray, I picked it up.
  [Read more]
  Published on April 16 2009 by J. W. Little

- An insult to Douglas Adams' memory
  The filmmaker's reverence for Adams' legacy? What kind of rubbish statement is that? As a loyal fan of Douglas Adams for more than a quarter of a century, I was appalled and...
  [Read more]
  Published on Aug 22 2006 by Daniel Jolley
Our Domain Adaptation Setting

Classification task
- Input space: $\mathcal{X} \subseteq \mathbb{R}^d$
- Labels: $\mathcal{Y} = \{0, 1, 2, \ldots, L\}$

Two different data distributions
- Source domain: $\mathcal{D}_S$
- Target domain: $\mathcal{D}_T$

A domain adaptation learning algorithm is provided with

- A labeled source sample $S = \{(x^s_i, y^s_i)\}_{i=1}^n \sim (\mathcal{D}_S)^n$.
- An unlabeled target sample $T = \{x^t_i\}_{i=1}^{n'} \sim (\mathcal{D}_T)^{n'}$.

The goal is to build a classifier $\eta: \mathcal{X} \rightarrow \mathcal{Y}$ with a low target risk

$$R_{\mathcal{D}_T}(\eta) \overset{\text{def}}{=} \Pr_{(x^t, y^t) \sim \mathcal{D}_T} [\eta(x^t) \neq y^t].$$
Domain Adaptation

Question
In which context can we adapt from source $\mathcal{D}_S$ to target $\mathcal{D}_T$?

Rough Answer
When domains $\mathcal{D}_S$ and $\mathcal{D}_T$ are $\ll$similar$\gg$.

Tool
Notion of “distance” $d(\mathcal{D}_S, \mathcal{D}_T)$ between domains.

Two approaches to conceive learning algorithms
1. Find a hypothesis $\eta \in \mathcal{H}$ such that $d_\eta(\mathcal{D}_S, \mathcal{D}_T)$ and $R_{\mathcal{D}_S}(\eta)$ are small.
2. Modify the representation of the examples:
   $\Rightarrow$ Find a function $h$ such that $d_{\mathcal{H}}(h(\mathcal{D}_S), h(\mathcal{D}_T))$ is small;
   and a $\eta \in \mathcal{H}$ such that $R_h(\mathcal{D}_S)(\eta)$ is small.
Divergence between source and target domains

**Definition (Ben David et al., 2006)**

Given two domain distributions $\mathcal{D}_S$ and $\mathcal{D}_T$, and a hypothesis class $\mathcal{H}$, the $\mathcal{H}$-divergence between $\mathcal{D}_S$ and $\mathcal{D}_T$ is

\[
d_{\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \overset{\text{def}}{=} 2 \sup_{\eta \in \mathcal{H}} \left| \Pr_{x^s \sim \mathcal{D}_S} [\eta(x^s) = 1] + \Pr_{x^t \sim \mathcal{D}_T} [\eta(x^t) = 0] - 1 \right|.
\]

The $\mathcal{H}$-divergence measures the ability of an hypothesis class $\mathcal{H}$ to **discriminate** between source $\mathcal{D}_S$ and target $\mathcal{D}_T$ distributions.
Bound on the target risk

**Theorem (Ben David et al., 2006)**

Let \( \mathcal{H} \) be a hypothesis class of VC-dimension \( d \). With probability \( 1 - \delta \) over the choice of samples \( S \sim (D_S)^n \) and \( T \sim (D_T)^n \), for every \( \eta \in \mathcal{H} \):

\[
R_{D_T}(\eta) \leq \hat{R}_S(\eta) + \frac{4}{n} \sqrt{d \log \frac{2en}{d} + \log \frac{4}{\delta}} + \hat{d}_\mathcal{H}(S, T) + \frac{4}{n^2} \sqrt{d \log \frac{2n}{d} + \log \frac{4}{\delta}} + \beta
\]

with \( \beta \geq \inf_{\eta^* \in \mathcal{H}} \left[ R_{D_S}(\eta^*) + R_{D_T}(\eta^*) \right] \).

Empirical risk on the source sample:

\[
\hat{R}_S(\eta) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I[\eta(x_i^s) \neq y_i^s].
\]

Empirical \( \mathcal{H} \)-divergence:

\[
\hat{d}_\mathcal{H}(S, T) \overset{\text{def}}{=} 2 \max_{\eta \in \mathcal{H}} \left[ \frac{1}{n} \sum_{i=1}^{n} I[\eta(x_i^s) = 1] + \frac{1}{n'} \sum_{i=1}^{n'} I[\eta(x_i^t) = 0] - 1 \right].
\]
Bound on the target risk

Theorem (Ben David et al., 2006)

Let $\mathcal{H}$ be a hypothesis class of VC-dimension $d$. With probability $1 - \delta$ over the choice of samples $S \sim (\mathcal{D}_S)^n$ and $T \sim (\mathcal{D}_T)^n$, for every $\eta \in \mathcal{H}$:

$$R_{\mathcal{D}_T}(\eta) \leq \hat{R}_S(\eta) + \frac{4}{n} \sqrt{d \log \frac{2e n}{d}} + \log \frac{4}{\delta} + \hat{d}_H(S, T) + \frac{4}{n^2} \sqrt{d \log \frac{2n}{d}} + \log \frac{4}{\delta} + \beta$$

with $\beta \geq \inf_{\eta^* \in \mathcal{H}} [R_{\mathcal{D}_S}(\eta^*) + R_{\mathcal{D}_T}(\eta^*)]$.

**Target risk** $R_{\mathcal{D}_T}(\eta)$ is low if, given $S$ and $T$, $\hat{R}_S(\eta)$ is small, i.e., $\eta \in \mathcal{H}$ is good on and $\hat{d}_H(S, T)$ is small, i.e., all $\eta' \in \mathcal{H}$ are bad on.
Let consider a neural network architecture with one hidden layer

\[ h(x) = \text{sigm}(Wx + b), \quad \text{and} \quad f(h(x)) = \text{softmax}(Vh(x) + c). \]

\[
\min_{W,V,b,c} \left[ \frac{1}{n} \sum_{i=1}^{n} -\log \left( f_y^n \left( h(x^s_i) \right) \right) \right].
\]

where \( f_y(h(x)) \) denotes the conditional probability that the neural network assigns \( x \) to class \( y \).

Given a source sample \( S = \{ (x^s_i, y^s_i) \}_{i=1}^{n} \sim (\mathcal{D}_S)^n \), \( W, b \)

1. Pick a \( x^s \in S \)
2. Update \( (V, c) \) towards \( f(h(x^s)) = y^s \)
3. Update \( (W, b) \) towards \( f(h(x^s)) = y^s \)

The hidden layer learns a representation \( h(\cdot) \) from which linear hypothesis \( f(\cdot) \) can classify source examples.
**Empirical $\mathcal{H}$-divergence**

\[
\hat{d}_\mathcal{H}(S, T) \overset{\text{def}}{=} 2 \max_{\eta \in \mathcal{H}} \left[ \frac{1}{n} \sum_{i=1}^{n} I[\eta(x_s^i) = 1] + \frac{1}{n'} \sum_{i=1}^{n'} I[\eta(x_t^i) = 0] - 1 \right].
\]

Given a representation output by the hidden layer $h(\cdot)$, we estimate the $\mathcal{H}$-divergence by

\[
\hat{d}_\mathcal{H}(h(S), h(T)) \approx 2 \max_{u,d} \left[ \frac{1}{n} \sum_{i=1}^{n} \log(o(h(x_s^i))) + \frac{1}{n'} \sum_{i=1}^{n'} \log(1 - o(h(x_t^i))) - 1 \right].
\]

where $o(h(x))$ is a logistic regressor that “tries” to detect if $x$ is from the **source domain** ($o(h(x)) > \frac{1}{2}$) or **target domain** ($o(h(x)) < \frac{1}{2}$) :

\[
o(h(x)) \overset{\text{def}}{=} \text{sigm}(u^\top h(x) + d).
\]
Domain-Adversarial Neural Network (DANN)

$$\min_{W, V, b, c} \left[ \frac{1}{n} \sum_{i=1}^{n} -\log \left( f_{y_i}^s (h(x_i^s)) \right) + \lambda \max_{u, d} \left( \frac{1}{n} \sum_{i=1}^{n} \log \left( o(h(x_i^s)) \right) + \frac{1}{n'} \sum_{i=1}^{n'} \log \left( 1 - o(h(x_i^t)) \right) \right) \right],$$

where $\lambda > 0$ weights the domain adaptation regularization term.

Given a source sample $S = \{(x_i^s, y_i^s)\}_{i=1}^{n} \sim (D_S)^n$, and a target sample $T = \{(x_i^t)\}_{i=1}^{n'} \sim (D_T)^{n'}$,

1. Pick a $x^s \in S$ and $x^t \in T$
2. Update $(V, c)$ towards $f(h(x^s)) = y^s$
3. Update $(W, b)$ towards $f(h(x^s)) = y^s$
4. Update $(u, d)$ towards $o(h(x^s)) = 1$
   and $o(h(x^t)) = 0$
5. Update $(W, b)$ towards $o(h(x^s)) = 0$
   and $o(h(x^t)) = 1$

DANN finds a representation $h(\cdot)$ that are good on $S$; but **un**able to **di**scriminate between $S$ and $T$. 

Pascal Germain (INRIA/Modal)  
Representation Learning Domain Adaptation  
November 7, 2017  
13 / 25
**Toy Dataset**

**Standard Neural Network (NN)**

Classification output: \( f(h(x)) \)

**Domain-Adversarial Neural Networks (DANN)**

Classification output: \( f(h(x)) \)

Representation PCA
Toy Dataset

Standard Neural Network (NN)

- Classification output: \( f(h(x)) \)
- Hidden Neurons

Domain-Adversarial Neural Networks (DANN)

- Classification output: \( f(h(x)) \)
- Hidden Neurons

\[ f_0(x) \quad f_1(x) \]
\[ h(x) \]
\[ x \]
\[ V, c \]
\[ W, b \]
\[ \alpha(x) \]
\[ u, d \]
Choosing the Hyperparameters

Model Selection by *Reverse Validation* (inspired by Zhong et al., 2010)

For each tuple of hyperparameters:

- Split $S$, $T$ into *training sets* $S'$, $T'$ and *validation sets* $S_V$, $T_V$.
- Learn classifier $\eta$ on (labeled) source $S'$ and (unlabeled) target $T'$.
- Learn **reverse classifier** $\eta_r$ on self-labeled $S'_r = \{ (x^t, \eta(x^t)) \}_{x^t \in T'}$ as source and unlabeled part of $S'$ as target.
- Compute the **reverse validation risk** $\hat{R}_{S_V}(\eta_r)$.

\[
T'_r = \{ (x^s_i) \}_{i=1}^{S'}
\]

\[
S_V = \{ (x^s_i, y^s_i) \}_{i=1}^{n}\]

\[
S'_r = \{ (x^t_i, \eta(x^t_i)) \}_{i=1}^{T'}
\]

\[
T_V = \{ (x^t_i) \}_{i=1}^{n'}
\]

\[
\hat{R}_{S_V}(\eta_r) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I[\eta(x^s_i) \neq y^s_i].
\]
**Input**: product review (bag of words)

**Output**: positive or negative rating.
<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>DANN</th>
<th>NN</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>books</td>
<td>dvd</td>
<td>.784</td>
<td>.790</td>
<td><strong>.799</strong></td>
</tr>
<tr>
<td>books</td>
<td>electronics</td>
<td>.733</td>
<td>.747</td>
<td><strong>.748</strong></td>
</tr>
<tr>
<td>books</td>
<td>kitchen</td>
<td>.779</td>
<td>.778</td>
<td>.769</td>
</tr>
<tr>
<td>dvd</td>
<td>books</td>
<td>.723</td>
<td>.720</td>
<td><strong>.743</strong></td>
</tr>
<tr>
<td>dvd</td>
<td>electronics</td>
<td>.754</td>
<td>.732</td>
<td>.748</td>
</tr>
<tr>
<td>dvd</td>
<td>kitchen</td>
<td>.783</td>
<td>.778</td>
<td>.746</td>
</tr>
<tr>
<td>electronics</td>
<td>books</td>
<td>.713</td>
<td>.709</td>
<td>.705</td>
</tr>
<tr>
<td>electronics</td>
<td>dvd</td>
<td>.738</td>
<td>.733</td>
<td>.726</td>
</tr>
<tr>
<td>electronics</td>
<td>kitchen</td>
<td><strong>.854</strong></td>
<td><strong>.854</strong></td>
<td>.847</td>
</tr>
<tr>
<td>kitchen</td>
<td>books</td>
<td>.709</td>
<td>.708</td>
<td>.707</td>
</tr>
<tr>
<td>kitchen</td>
<td>dvd</td>
<td>.740</td>
<td>.739</td>
<td>.736</td>
</tr>
<tr>
<td>kitchen</td>
<td>electronics</td>
<td><strong>.843</strong></td>
<td>.841</td>
<td>.842</td>
</tr>
</tbody>
</table>
JMLR 2016: **Domain-Adversarial Neural Networks.**
by Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand and Lempitsky

http://jmlr.org/papers/v17/15-239.html
Gradient Reversal Layer

Implemented in Caffe Deep Learning Package (Jia et al. 2014):

\[ R(x) = x, \quad \frac{dR}{dx} = -I. \]
<table>
<thead>
<tr>
<th>Method</th>
<th>Source Target</th>
<th>MNIST</th>
<th>Syn Numbers</th>
<th>SVHN</th>
<th>Syn Signs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Source only</td>
<td>.5225</td>
<td>.8674</td>
<td>.5490</td>
<td>.7900</td>
</tr>
<tr>
<td></td>
<td>SA (Fernando et al., 2013)</td>
<td>.5690</td>
<td>.8644</td>
<td>.5932</td>
<td>.8165</td>
</tr>
<tr>
<td></td>
<td>DANN</td>
<td><strong>.7666</strong></td>
<td><strong>.9109</strong></td>
<td><strong>.7385</strong></td>
<td><strong>.8865</strong></td>
</tr>
<tr>
<td></td>
<td>Train on target</td>
<td>.9596</td>
<td>.9220</td>
<td>.9942</td>
<td>.9980</td>
</tr>
</tbody>
</table>
### Office Dataset

**Images from three domains**: Amazon, DSLR camera, and Webcam

**31 labels**: chair, cup, laptop, keyboard, ...

<table>
<thead>
<tr>
<th>Method</th>
<th>Source Target</th>
<th>Amazon Webcam</th>
<th>DSLR Webcam</th>
<th>Webcam DSLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFK (PLS, PCA) (Gong et al. 2012)</td>
<td></td>
<td>.197</td>
<td>.497</td>
<td>.6631</td>
</tr>
<tr>
<td>SA (Fernando et al., 2013)</td>
<td></td>
<td>.450</td>
<td>.648</td>
<td>.699</td>
</tr>
<tr>
<td>DLID (Chopra et al., 2013)</td>
<td></td>
<td>.519</td>
<td>.782</td>
<td>.899</td>
</tr>
<tr>
<td>DDC (Tzeng et al., 2014)</td>
<td></td>
<td>.618</td>
<td>.950</td>
<td>.985</td>
</tr>
<tr>
<td>DAN (Long and Wang, 2015)</td>
<td></td>
<td>.685</td>
<td>.960</td>
<td>.990</td>
</tr>
<tr>
<td>Source only</td>
<td></td>
<td>.642</td>
<td>.961</td>
<td>.978</td>
</tr>
<tr>
<td>DANN</td>
<td></td>
<td><strong>.730</strong></td>
<td><strong>.964</strong></td>
<td><strong>.992</strong></td>
</tr>
</tbody>
</table>
Derivative works

- Domain Adversarial Training for **Accented Speech Recognition**, Sun et al., 2017
- Adversarial Multi-task Learning for Text Classification, Liu et al., 2017
- Domain-Adversarial Neural Networks to Address the Appearance Variability of **Histopathology Images**, Lafarge et al., 2017
- Unsupervised Domain Adaptation in **Brain Lesion Segmentation** with Adversarial Networks, Kamnitsas et al., 2016
- Variational Adversarial Deep Domain Adaptation for **Health Care Time Series Analysis**, Purushotham et al., 2016
- **Predicting Sales** from the Language of Product Descriptions, Pryzant et al., 2017
- No More Discrimination : Cross City Adaptation of **Road Scene Segmenters**, Chen et al., 2017
- Using Simulation and Domain Adaptation to Improve Efficiency of Deep **Robotic Grasping**, Bousmalis et al., 2017
- Multi-task Domain Adaptation for Deep Learning of **Instance Grasping from Simulation**, Fang et al., 2017
- Adversarial Multi-Criteria Learning for **Chinese Word Segmentation**, Chen et al., 2017
- Adversarial Domain Adaptation for **Identifying Phase Transitions**, Huembeli et al., 2017
- Use the domain regression output to estimate the target/source density

\[
\frac{p(x|D_T)}{p(x|D_S)} \approx \frac{1 - o(x)}{o(x)}.
\]

- Use the classification output to enforce large margin on target samples
We learn a new representation that is

1. accurate on the source domain; but
2. unable to discriminate between source and target domains.

Our method is:

- Directly based on the seminal theory of domain adaptation of Ben-David et al. (2006).
- Easy to implement in any neural network architectures.
- Achieving state-of-the-art results on several benchmarks.