Marginalized Stacked Denoising Autoencoders for Domain Adaptation

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha
Domain Adaptation

Training:
Source Domain
Sufficient data

Testing:
Target Domain
Insufficient data

$P_S(Y|X)$  $P_h(Y|X)$  $P_T(Y|X)$
Showcase application: Sentiment Analysis

predict sentiment (star rating) from review text

[Blitzer et al., 2006]

Most Helpful Customer Reviews

29,534 of 29,875 people found the following review helpful:

⭐⭐⭐⭐⭐ Kindle vs. Nook (updated 6/2/2011), August 28, 2010
By Ron Cronovich "Ron" (Kenosha, WI) - See all my reviews

Top 1000 review

Amazon Verified Purchase (What's this?)
This review is from: Kindle 3G, Free 3G + Wi-Fi, 3G Works Globally, 6" Display with New E Ink Pearl Technology (Electronics)

When I wrote this review in August 2010, there was only one Nook, which is now called "Nook First Edition." It continues to be available, but there are now two new Nooks. The Nook Color was introduced last fall - it's basically a tablet computer, and runs the Android software that is popular on many smartphones nowadays. It's twice as heavy and costs twice as much as a Kindle, but compared to other tablet computers, it is a very good value.

And now (early June 2011), a new e-ink based Nook is coming out. It's called the "Nook Simple Touch." It is just now starting to ship, so obviously I don't have one and can't tell you anything about it that you can't learn by reading online reviews. But the reviews are very favorable, so if you're considering a Kindle, you should take a look at the new Nook Simple Touch, too.

But the Kindle is nonetheless still a compelling option. It's a mature product, very well designed and easy to use, performance is very zippy, it's competitively priced, and no e-ink based reader has a better, more readable display than the Kindle, not even the new Nook Simple Touch. Also, the Kindle universe is quite extensive: the Kindle store is great and has many thousands of free e-books as well as good deals on most other e-books, and once purchased, you can read your Kindle books on nearly any device you own (computer, phone, tablet), not just your Kindle. And there are tons of great cases and other accessories for the Kindle.

So, while my review compares the Kindle to the older Nook, I'll leave it here because it has a ton of information about the Kindle, a great e-reader that deserves your attention, and because the original Nook continues to be available. That said, I urge you to NOT buy the original Nook. It was a respectable e-reader when it came out in 2009, and still had some value when I wrote about it in August 2010, but it is clearly inferior by today's standards.

------------- my original review ---------------

If you're trying to choose between a Nook and a Kindle, perhaps I can help. My wife and I have owned a Nook (the original one), a Kindle 2, and a Kindle DX. When Amazon announced the Kindle 3 this summer, we pre-ordered two Kindle 3's: the wi-fi only model in graphite, and the wi-fi + 3G model in white. They arrived in late August and we have used them very regularly since then. For us, Kindle is better than Nook, but Nook is a good device with its own advantages that I will discuss below. I'll end this review with a few words about the Nook Color.
Same Task, Different Distribution

Naïve approach: train on source, test on target

Source domain: Books
(Test Error: 13%)

Target domain: Kitchen Appliances
(Test Error: 24%)

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha
Marginalized Stacked Denoising Autoencoders
I read 2-3 books a week, and this is without a doubt my favorite of this year. A beautiful novel by Afghan-American Khaled Hosseini that ranks among the best-written and provocative stories of the year so far. This unusually eloquent story is also about the fragile relationship.

This unit makes the best coffee I've had in a home. It is my favorite unit. It makes excellent and HOT coffee. The carafe is solidly constructed and fits securely in square body. The timer is easy to program, as is the clock.

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha
Marginalized Stacked Denoising Autoencoders
Learning Joint Feature Representation

[Blitzer et al., 2007; Satpal and Sarawagi, 2007; Pan et al., 2010]
Denoising Autoencoders (DA)

[Vincent et al., 2008; Glorot et al., 2011]

reviews from source or target

bag of words

favorite
best-written
eloquent
energy-efficient
solidly-constructed

reconstruction

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha

Marginalized Stacked Denoising Autoencoders
Denoising Autoencoder (DA)

**Vincent et al., 2008**

- favorite
- best-written
- eloquent
- energy-efficient
- solidly-constructed

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha

Marginalized Stacked Denoising Autoencoders

---

\[
\min_W \| x - g(W^T f(W\tilde{x})) \|^2
\]
Stacked Denoising Autoencoder (SDA)

[Vincent et al., 2008]
SDA for Feature Learning

\[ \text{x} \rightarrow h_1 \rightarrow h_2 \rightarrow \text{new representation} \]

\[ [\text{Glorot et al., 2011}] \]

linear classifier (SVM)
Accurate but Slow

- SDAs generate robust features for domain adaptation
- Pre-training of SDAs requires (stochastic) gradient descent, slow for large scale dataset
- dense-matrix GPU implementation [Bergstra, et al., 2010]
- reconstruction sampling [Dauphin et al., 2011]

Hyper-parameters (learning rate, number of epochs, noise ratio, mini-batch size, network structure, etc.) [Bergstra and Bengio, 2011]

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha
Marginalized Stacked Denoising Autoencoders
Research Goal

![Graph showing the trade-off between training time and transfer ratio.](image)

**Keep the accuracy, improve the speed**

- **Bag-of-words** (0 secs)
- **PCA** (~3 mins)
- **SCL** (47 secs) [Blitzer et al., 2007]
- **CODA** (~25 mins) [Chen et al., 2011]
- **SDA** (~5 hours) [Glorot et al., 2011]

**Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha**
marginalized Denoising Autoencoder (mDA)

Convexity

\[
\ell_{DA} = \sum_{i=1}^{n} \| x_i - g(W^T f(W \tilde{x})) \|^2
\]

\[
\ell = \sum_{i=1}^{n} \| x_i - W \tilde{x}_i \|^2
\]

local minima

global minima

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha
marginalized Denoising Autoencoder (mDA)
Closed-form solution

\[ W = \left( \sum_{i=1}^{n} x_i \tilde{x}_i^\top \right) \left( \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^\top \right)^{-1} \]

\[ \ell = \sum_{i=1}^{n} \| x_i - W \tilde{x}_i \|^2 \]

Any initialization

Global minima
marginalized Denoising Autoencoder (mDA)

Multiple corruptions

\[
W = \left( \sum_{i=1}^{n} x_i \tilde{x}_i^T \right) \left( \sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T \right)^{-1}
\]

\[
\ell = \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \| x_i - W \tilde{x}_{i,j} \|^2
\]

\[
W = \left( \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} x_{i,j} \tilde{x}_{i,j}^T \right) \left( \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \tilde{x}_{i,j} \tilde{x}_{i,j}^T \right)^{-1}
\]
marginalized Denoising Autoencoder (mDA)

Marginalized corruption

\[
\tilde{x} \xrightarrow{W} z
\]

\[
W = \left( \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} x_i \tilde{x}_{i,j} \right) \left( \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \tilde{x}_{i,j} \tilde{x}_{i,j}^\top \right)^{-1}
\]
marginalized Denoising Autoencoder (mDA)

Marginalized corruption

\[ W = \left( \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} x_i \tilde{x}_{i,j}^{T} \right) \left( \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \tilde{x}_{i,j} \tilde{x}_{i,j}^{T} \right)^{-1} \]

\[ W^* = \left( \sum_{i=1}^{n} \mathbb{E}[x_i \tilde{x}_{i}^{T}] \right) \left( \sum_{i=1}^{n} \mathbb{E}[\tilde{x}_{i} \tilde{x}_{i}^{T}] \right)^{-1} \]

- closed form!
- corruption marginalized out!

\[ P_{\alpha,\beta} = S_{\alpha\beta} q_{\beta} \]

\[ Q_{\alpha,\beta} = S_{\alpha\beta} q_{\alpha} q_{\beta} \]
marginalized Stacked Denoising Autoencoder (mSDA)
mSDA for Feature Learning

\[ x \rightarrow h_1 \rightarrow h_2 \rightarrow \text{new representation} \rightarrow \text{linear classifier (SVM)} \]
20 lines of Matlab

**Algorithm 1** mDA in MATLAB™.

```matlab
function [W,h]=mDA(X,p);
X=[X;ones(1,size(X,2))];
d=size(X,1);
q=ones(d,1).*(1-p); q(end)=1;
S=X*X';
Q=S.*(q*q');
Q(1:d+1:end)=q.*diag(S);
P=S.*repmat(q,1,d);
W=((Q+1e-5*eye(d))\P(:,1:end-1))';
h=W*X;
```

**Algorithm 2** mSDA in MATLAB™.

```matlab
function [Ws,hs]=mSDA(X,p,l);
[d,n]=size(X);
Ws=zeros(d,d+l,1);
hs=zeros(d,d+l);
hs(:,:,1)=X;
for t=1:l
    Ws(:,:,t)=mDA(hs(:,:,t),p);
    hst=Ws(:,:,t)*[hs(:,:,t);ones(1,n)];
    hs(:,:,t+1)=tanh(hst);
end;
```
Results: Amazon review dataset

[Blitzer et al., 2006]
Experimental Results

\[ \text{transfer loss} = \text{error}(D \rightarrow B) - \text{error}(E \rightarrow B) \]

4 domains, 12 tasks
\[ d = 5,000 \]
\[ n = 27,677 \]

Baseline
PCA
SCL (Blitzer et. al., 2007)
CODA (Chen et. al., 2011)
SDA (Glorot et. al., 2011)
mSDA (l=5)

Experimental Results

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha
Marginalized Stacked Denoising Autoencoders
Transfer ratio vs. training time

4 domains, 12 tasks
\(d = 5,000\)
\(n = 27,677\)

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha
Marginalized Stacked Denoising Autoencoders

Sunday, October 13, 13
Large data set

20 domains, 380 tasks
\( d = 5,000 \)
\( n = 339,675 \)
High Dimensional Data

4 domains, 12 tasks
\(d = 5,000 - 40,000\)
\(n = 27,677\)

\[10^2\] \[10^3\] \[10^4\] \[10^5\]

\(d = 5,000\)
\(d = 10,000\)
\(d = 20,000\)
\(d = 40,000\)
\(d = 5,000\)

Transfer Ratio

\(~3\) mins \(\rightarrow\) faster

\(~19\) hours

\(400x\) speedup

\(\text{mSDA}\)

\(\text{SDA (Glorot et. al., 2011)}\)

Minmin Chen, Zhixiang (Eddie) Xu, Kilian Q. Weinberger, Fei Sha

Marginalized Stacked Denoising Autoencoders
Conclusion

- **marginalized Stacked Denoising Autoencoder (mSDA)***
  - marginalizes out corruption
  - keeps high accuracy of SDAs but is orders of magnitudes faster
- **Optimization:**
  - is layer-wise convex
  - has layer-wise closed form solutions
  - is easy to implement
Conclusion

- **marginalized Stacked Denoising Autoencoder (mSDA)**
- marginalizes out corruption
- keeps high accuracy of SDAs but is orders of magnitudes faster

**Optimization:**
- is layer-wise convex
- has layer-wise closed form solutions
- is easy to implement
Conclusion

- **marginalized Stacked Denoising Autoencoder (mSDA)**
  - marginalizes out corruption
  - keeps high accuracy of SDAs but is orders of magnitudes faster

Optimization:
- is layer-wise convex
- has layer-wise closed form solutions
- is easy to implement
Conclusion

- **marginalized Stacked Denoising Autoencoder (mSDA)**
  - marginalizes out corruption
  - keeps high accuracy of SDAs but is orders of magnitudes faster

- **Optimization:**
  - is layer-wise convex
  - has layer-wise closed form solutions
  - is easy to implement
Conclusion

- **marginalized Stacked Denoising Autoencoder (mSDA)**
  - marginalizes out corruption
  - keeps high accuracy of SDAs but is orders of magnitudes faster

**Optimization:**
- is layer-wise convex
- **has layer-wise closed form solutions**
- is easy to implement
Conclusion

- **marginalized Stacked Denoising Autoencoder (mSDA)**
  - marginalizes out corruption
  - keeps high accuracy of SDAs but is orders of magnitudes faster
- **Optimization:**
  - is layer-wise convex
  - has layer-wise closed form solutions
  - is easy to implement

**Algorithm 1** mDA in MATLAB™

```matlab
function [W,h]=mDA(X,p);
    X=[X;ones(1,size(X,2))];
    d=size(X,1);
    q=ones(d,1).*(1-p); q(end)=1;
    S=X'*X;
    Q=S.*(q'*q);
    Q(1:d+1:end)=q.*diag(S);
    P=S.*repmat(q,1,d);
    W=((Q+1e-5*eye(d))\P(:,:,1:end-1))';
    h=W*X;
end
```

**Algorithm 2** mSDA in MATLAB™

```matlab
function [s,hs]=mSDA(D)
    [d,n]=size(A);
    Ws=zeros(d,d+1,:);
    hs=zeros(d,d,1+1);
    hs(:,:,1)=X;
    for t=1:l
        Ws(:,:,t)=mDA(hs(:,:,t),p);
        hst=Ws(:,:,t)*[hs(:,:,t);ones(1,n)];
        hs(:,:,t+1)=tanh(hst);
    end
end
```
Thank you!

Questions?