Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

Heterogeneous Transfer Learning with RBMs

Bin Wei
University of Rochester
everfool7@gmail.com

Christopher Pal
Ecole Polytechnique de Montreal
christopher.pal@polymtl.ca

Abstract

A common approach in machine learning is to use a large amount of labeled data to train a model. Usually this model can then only be used to classify data in the same feature space. However, labeled data is often expensive to obtain. A number of strategies have been developed by the machine learning community in recent years to address this problem, including: semi-supervised learning, domain adaptation, multi-task learning, and self-taught learning. While training data and test may have different distributions, they must remain in the same feature set. Furthermore, all the above methods work in the same feature space. In this paper, we consider an extreme case of transfer learning called heterogeneous transfer learning - where the feature spaces of the source task and the target tasks are disjoint. Previous approaches mostly fall in the multi-view learning category, where co-occurrence data from both feature spaces is required. We generalize the previous work on cross-lingual adaptation and propose a multi-task strategy for the task. We also propose the use of a restricted Boltzmann machine (RBM), a special type of probabilistic graphical models, as an implementation. We present experiments on two tasks: action recognition and cross-lingual sentiment classification.

Introduction

In machine learning we often come across scenarios where we have plenty of labeled data in an “older” feature space, but we are short of examples in the new feature space. For example, there are a lot of annotated corpora in English, but there are usually far fewer in other languages or even none for some small languages. We are thus motivated to transfer knowledge between different feature spaces. This problem has been characterized as “translated learning” (Dai et al. 2008) or “heterogeneous transfer learning” (Yang et al. 2008).

To address the problem introduced by different feature spaces, researchers have studied the problem of multi-view learning (Blum and Mitchell 1998). In multi-view learning, each instance has multiple views in different feature spaces. Yet this approach is inapplicable in many real life applications where we typically have training data in a source feature space, the test data in a totally different target feature space and there may be no correspondence between instances in these spaces. Heterogeneous transfer learning thus relaxes many constraints on the problem compared to other types of learning and allows a wide variety of more flexible set-ups for applications.

Dai et al. (2008) proposed to learn a feature translator $p(x_t|x_s)$ and optimize the model under the risk minimization framework. To estimate the feature translator it is required to have some co-occurrence data across the two feature spaces. The authors performed experiments on two tasks: cross-language classification and text-aided image classification. For the cross-language task the co-occurrence data is extracted from electronic dictionaries while for image classification annotated images are extracted from social websites and search engines.

The idea of Dai et al. is straightforward, yet one may question the necessity of building such a feature level translator. After all, in most real world applications we have a very high dimension feature space and often only a small number of features are useful. Further, we are often not really interested in learning how to generate pixel intensities from words and what we really need is a classifier in the target space.

Yang et al. (2008) also studied the text aided image processing problem. Instead of going through the indirect translator, the authors proposed to directly model the relations between features from different spaces and the target label we want to predict. A modified PLSA model is designed to model the two co-occurrence matrices: the one between image clusters and visual features and the one between text annotations and visual features. Their co-occurrence data was obtained from social websites like Flickr.

The task of heterogeneous transfer learning is difficult because the feature spaces are non-overlapping. Unlike other learning scenarios, one has no natural bridge between the two spaces. While older work has relied on co-occurrence data, some recent effort (Prettenhofer and Stein 2010; Wei and Pal 2010) has been directed towards adapting a domain adaptation algorithm called structural correspondence learning (SCL) (Blitzer, Dredze, and Pereira 2007; Blitzer, McDonald, and Pereira 2006) to perform cross-lingual document classification and obtained positive results.

The SCL algorithm was initially designed for semi-supervised leaning by learning multiple related tasks together (Ando and Zhang 2005). This inspired us to use a similar idea for heterogeneous transfer learning. In our work here, we will connect the two feature spaces by mapping them to a common representation or latent space. To obtain such a mapping, we consider how each feature functions in a set of selected tasks. The assumption we make here is that if we have a set of tasks semantically related to our target task, and if some features behave similarly in these related tasks, they should also have similar roles in the target problem (even they are from different feature spaces). Capturing these similarities will enable us to connect the two feature spaces to transfer knowledge. Note that different from Dai et al. (2008), we do not ask for a full feature-level translator, what we seek here is a shared representation specific to our target task. We call this setting multi-task heterogeneous transfer learning. Any specific algorithm in this setting will aim at finding a representation that works well for the set of related task. Specifically in this paper, we propose an example approach using restricted Boltzmann machines (RBMs).

The following section describes how we can use RBMs for multitask heterogeneous transfer learning in a general sense. We then present the exact RBM model we use for two experiments: text-aided simple activity recognition cross-lingual sentiment classification in the experiment section.

RBMs for multitask heterogeneous transfer learning

Restricted Boltzmann machines (RBMs) have long been used in problems like feature selection. Yet they can also be used as general purpose classifiers and been shown to achieve state of the art performance (Larochelle and Bengio 2008). They have also been used as topic models using the term "harmonium" model(s) in tasks like video topic identifications (Yang et al. 2007) and information retrieval (Welling, Rosen-Zvi, and Hinton 2004). One of our goals in this paper is to show that although simple, RBMs can be a useful tool for transfer learning.

Problem formulation

Assume we have a learning task T_T, and we have two feature spaces X_s and X_t. We denote $X_s = \{x_{si}\}^N$ as the source feature space, and $X_t = \{x_{tk}\}^M$ the target feature space. We have $X_s \cap X_t = \emptyset$.

We also have a set of related/auxiliary tasks $\{Ta_i\}$. And again for each auxiliary task Ta_i, we observe instances from both X_s and X_t. We have training instances for the auxiliary problems in the two spaces as $(\{x_{si}\}, Ta_i)$ and $(\{x_{tk}\}, Ta_i)$ but they don't need to be multi-view data. Our goal is to use these auxiliary training instances plus some training instances $(\{x_{si}\}, T_T)$ for the target problem T_T in the source feature space to make a better prediction of T_T given $\{x_{tk}\}$.

To transfer knowledge between the two feature spaces, we rely on the semantic correspondence among the auxiliary tasks. In contrast with the usual set up in multi-task learning, we are now working on two different feature spaces. But, the basic idea is still the same: we want to find a good low dimensional representation shared by multiple auxiliary tasks. As an example, consider the small activity recognition problem we study here where we have text features as X_s and video features as X_t. For the 4 activities in Figure 1, we have some corresponding texts (dictionary glosses) as in Table 1. Now, if in the text domain, the words "hand" and "foot" are identified as important keywords, we can then expect the action type "running" to be more closely related to "walking" compared with "waving" and "clapping". We can make this guess purely based on the text features and the video features of the auxiliary types.

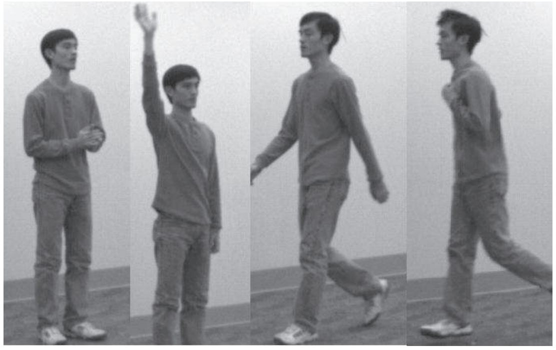

Figure 1: Simple activities, from left to right: clapping, waving, walking, running

Table 1: Simple activities, text data

Activity	Text
Clapping	hit your hand together that make a sound
Waving	move the hand or arm from side to side
Walking	to move or travel on leg and foot alternately
Running	go at fast pace to move rapidly on foot

We use the RBMs to model the intuition above. An RBM is a kind of bipartie graphical models where different observed variables (features and labels) are connected through a set of latent variables $H = \{h_j\}$, which will act as the low dimensional representation shared by the two different feature spaces.

To use RBMs for heterogeneous transfer learning, we first need to represent the two feature spaces, X_s and X_t, and the label variables Y_t for T_T and $\{Ta_i\}$.

We follow the approach proposed in Welling, Rosen-Zvi, and Hinton by starting with an independent model consisting of distributions from the exponential family, then connecting them with a hidden layer H. As an example, suppose we have chosen the suitable marginal distributions for X_s and H as :

$$p(X) = R(X)\exp\{\sum_a \theta_a f_a(X) - A(\{\theta_a\})\}$$

$$p(H) = S(H)\exp\{\sum_b \lambda_b g_b(H) - B(\{\lambda_b\})\}$$

where we have feature functions $f_a(X)$ and $g_b(H)$, parameters θ_a and λ_b, normalization factors $A(\{\theta_a\})$ and $B(\{\lambda_b\})$, and the standard additional functions: $R(X)$ and $S(H)$. We then alter these initial models by coupling them with a matrix W to obtain:

$$\begin{aligned} P(X,H) \quad \propto \quad & \exp\{\sum_a \theta_a f_a(X) + \sum_b \lambda_b g_b(H) + \\ & \sum_{a,b} W_{a,b} f_a(X) g_b(H)\} \end{aligned} \tag{1}$$

Importantly, the new joint model is not a member of the exponential family, but the conditional distributions are simply versions of exponential family distributions that are shifted by a matrix operation applied to their sufficient statistics. This fact leads to fast inference procedures which are often critical for practical run-times during learning.

The approach of coupling X and H in (1) can be repeated for coupling Y and H as discussed in (Yang et al. 2007) and (Wang, Pal, and McCallum 2007). We will use a matrix U to encode the Y and H coupling. One can then view the next step of our approach for our heterogeneous transfer learning as the creation of two discriminative RBMs, one for the *marginal conditional probability* $P(Y|X_s)$ using only X_s (with coupling matrix W) and one for $P(Y|X_t)$, using only X_t (using coupling matrix V), followed by a form of parameter tying for the U matrices for the label to hidden variable interactions. In other words, we couple H and the label variables to obtain an RBM for the source feature space and repeat the same steps and obtain another RBM for the target feature space. Since the two RBMs are dealing with the same T and $\{Ta_i\}$, we require that they share the same coupling matrix from H to the label variables. The same idea could also be implemented via a single discriminative RBM model with the constraint that unobserved features between the source and target feature sets have no impact on the hidden representation. We present this type of RBM in figure 2.

The two RBMs have their own mapping from the features to the hidden layer H respectively but must share the same mapping from H to Y. Our objective function is the combination of conditional log likelihood in both spaces:

$$L = \alpha \sum log(p(\{y\}|\{x_{si}\}) + (1-\alpha) \sum log(p(\{y\}|\{x_{tk}\})), \tag{2}$$

where the sums are over the training examples and we have omitted subscripts to simplify the equation. Note that as we introduce the hidden layer H, the features X_s and X_t are no longer directly connected to each other as in previous work, and this allow us to relax requirements on the co-occurrence data. Meanwhile, the connection between features is now built on their relations to the auxiliary problems. Features, whether from the same or different feature spaces, will be grouped together in the hidden layer if they behave similarly regarding to the auxiliary problems. Again, consider the activity example, the word "foot" and the image part of foot will be connected as they are both good indicators of action types "walking" and "running" but not for "waving" and "clapping". The latent variables H thus encode the correspondences between the two feature spaces.

Furthermore, the exponential family RBM framework allows one to easily incorporate different types of features, whether they are continuous or discrete, into our models, as long as they can be modeled as members of the exponential family. This property is especially helpful in the heterogeneous transfer learning case as we may have bags of words in one feature space, and continuous pixel intensities in the other.

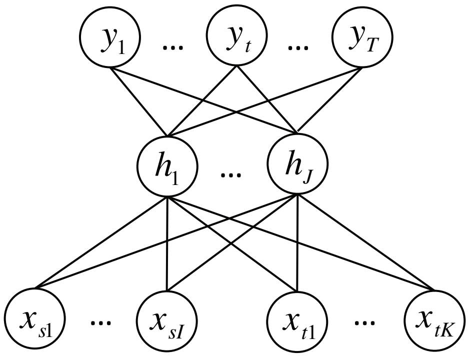

Figure 2: An RBM for multi-task heterogeneous transfer learning

Relation with SCL

The SCL approach of Blitzer, McDonald, and Pereira (2006) can also be viewed as a way to implement our multi-task heterogeneous transfer learning strategy. In both SCL and RBMs, we introduce a set of hidden variables and a set of auxiliary tasks. While SCL is a discriminative approach, RBMs are probabilistic models which can be formulated as joint or conditional (discriminative) models. One advantage of using probabilistic models is the simplicity of incorporating different kinds of variables as mentioned above. Furthermore, after we obtain the mappings from the training process, the SCL approach will need to deterministically calculate the value of hidden variables. On the other hand, for RBMs we don't have to explicitly infer the values for the hidden layer, we can simply integrate out the hidden variables and perform classification based on the observed features.

Experiment

A toy example: Text-aided simple human activity recognition

In this experiment, we consider how to incorporate text information with the 'old" activities to recognize new human activities. This problem is derived from the application where a user wants to identify a new action which he has not seen before. Our goal is to show that with the help of texts, we can identify a new action more accurately. In this experiment, the source feature space is the text while the target space is the video.

Data set We select 8 activities from the 13-action UIUC dataset [1] reported in Tran and Sorokin (2008). The 8 action types are walking, running, jumping, jumping jacks, waving, clapping, crawling, and turning. We use the same video features as in Tran and Sorokin (2008) where the target feature space $X_t = \{x_{tk}\}$ contain 286 dimension continuous features. We also re-sample the videos by the ratio 1:4: choosing 1 from every 4 consecutive frames. The number of frames varies from 13 to 45. There are 38 videos for each actions and the total size is 312.

For the text data, we extract texts mainly from different on-line dictionaries (cambridge [2], webster [3], merrian-webster [4],wiki [5], free-dict [6], msn [7]) We simply extract the first sense for each action except for the "jumping jacks" which is mainly the second sense in most of the dictionaries. We extract around 20 documents for each action type and after preliminary stop words removal and data cleaning we have a vocabulary about 500 words. X_s contain simple unigram features and modeled as binary features.

Model For a video clip, we have one set of hidden layer variables $\{h_{fj}\}$ for each frame f. We model $\{h_{fj}\}$ and$\{x_{tfk}\}$ with normal distribution and discrete distribution for $\{y\}$. The joint likelihood after coupling is as follows:

$$\begin{aligned} p(\{x_{tfk}\}, \{h_{fj}\},\{y\}) \propto & \\ & \exp\Bigg\{ \sum_t \sum_j \Bigg(\sum_k V_{jk} h_{fj} x_{tfk} \\ & + \sum_l U_{jl} h_{fj} y - \sum_k \frac{x_{tfk}^2}{2} - \sum_j \frac{h_{fj}^2}{2} \Bigg) \Bigg\} \end{aligned}$$

We optimize the conditional likelihood:

$$p(\{y\}|\{x_{tfk}\}) \propto \exp\Bigg\{ \sum_t \sum_j \Bigg(\sum_k V_{jk} x_{tfk} + \sum_l U_{jl} y \Bigg)^2 \Bigg\}$$

Similarly, for the text features $\{x_{si}\}$, we have:

$$\begin{aligned} p(\{x_{si}\}, \{h_j\}, \{y\}) \propto & \\ & \exp\Bigg\{ \sum_j \Bigg(\sum_i W_{ji} h_j x_{si} + \sum_l U_{jl} h_j y - \sum_j \frac{h_j^2}{2} \Bigg) \Bigg\}, \end{aligned}$$

$$p(\{y\}|\{x_{si}\}) \propto \exp\Bigg\{ \sum_j \Bigg(\sum_i W_{ji} x_{si} + \sum_l U_{jl} y \Bigg)^2 \Bigg\}$$

We need to estimate the three coupling matrices: W, U,and V. The corresponding gradient updates are listed below:

$$\begin{aligned} \partial V_{jk} &= E_m(x_{tk} U_{jl} y) - E_{p(y|\{x_{tk}\})}(x_{tk} U_{jy} y) \\ \partial W_{ji} &= E_m(U_{jl} y) - E_{p(y|\{x_{si}\})}(U_{jy} y) \\ \mathrm{d}U_{x_s} &= E_m\Bigg(\sum_i W_{ji} x_{si} + \sum_l U_{jl} y \Bigg) - \\ & \quad E_{p(y|\{x_{si}\})}\Bigg(\sum_i W_{ji} x_{si} + \sum_l U_{jy} y \Bigg) \\ \mathrm{d}U_{x_t} &= E_m\Bigg(\sum_k V_{jk} x_{tk} + \sum_l U_{jl} y \Bigg) - \\ & \quad E_{p(y|\{x_{tk}\})}\Bigg(\sum_k V_{ji} x_{tk} + \sum_l U_{jy} y \Bigg) \\ \partial U_{jl} &= \alpha \mathrm{d}U_{x_s} - (1-\alpha) \mathrm{d}U_{x_t} \end{aligned}$$

Here E_m stands for empirical expectation, and α is the parameter for trade-off between the two feature spaces.

Experiment design and results The task we perform in the experiment is to identify new actions that has not been seen before. Specifically, for each action type, we draw randomly one video clip from each of the remaining seven actions as the supervised data set in the video feature space. That is, in each experiment, we will have training video clips for 7 known actions, yet we have no videos for the target action. In the test stage, we will be given a mixture of videos from both the 7 seen action types and the unseen target action. The final model not only needs to distinguish the 7 actions it has seen, but also needs to identify the unseen target action from the other actions. This setting is denoted as "video labeling with rejection" (Tran and Sorokin 2008).

To solve this problem purely in the video space, we use the 1-NNMR approach taken in Tran and Sorokin (2008). For each frame, 1-NNMR looks for its closest neighbor frame within a predefined radius R, if it finds one, it uses the type of the frame otherwise reports "new". Each frame votes to get the label of the whole video sequence.

We compare this with an approach using our RBM constructed as outlined above to incorporate the text information. As we discussed earlier, we build two RBMs, one for the video space which uses the same training video dataset as 1-NNMR with the target action missing. Meanwhile, we also construct an RBM on the text space with all 8 actions' training texts.

In a word, for both the RBM model and the 1-NNMR, there is no video clips of the new action provided. But compared to 1-NNMR, our RBM model will have the extra information from texts. In both approaches, we both force every frame of the same clip to have the same label, so we will have dozens of training instances for each action in the video space. By comparing the performance of 1-NNMR and the RBM, we should be able to tell whether our RBM model succeeds in connecting the video and text features.

In the test stage, when given a video clip from the unseen target action, the 1-NNMR is supposed to give the label "new", while the RBM should simply label it to the target category as it has labeled texts for the target action. We mea-

[1] http://vision.cs.uiuc.edu/projects/activity/

[2] http://dictionary.cambridge.org/dictionary/

[3] http://www.webster-dictionary.org/

[4] http://www.merriam-webster.com/

[5] http://simple.wikipedia.org/wiki/

[6] http://www.thefreedictionary.com/

[7] http://encarta.msn.com/encnet/features/dictionary/

Missing Actions	**1-NNMR**	**RBM**	**% Error Reduction**
jumping jack	50.12	48.71	2.81
crawl	66.81	60.30	9.74
jump	60.76	56.33	7.29
clap	53.69	44.31	17.47
run	62.89	55.01	12.53
turn	63.06	58.73	6.87
walk	57.45	49.02	14.67
wave	54.72	48.13	12.04
average	*58.69*	*52.57*	*10.43*

Table 2: Results for activity recognition, in the number of misclassified video clips

sure the performance by simply counting how many mislabeled (“new” or one of the 7 training types) video clips in the whole test set. For each set of experiment, we sample 20 different training sets and reports the average. The results are shown in Table 2.

As we can see, results in the table show some positive signs of our RBM method in capturing the semantic information, the hybrid text-video method outperforms the pure video-based approach by reducing the average error by around 10%. The extra knowledge from the text space indeed helps improve the performance. The improvement benefits some actions more than others, as one might expect. One explanation for this phenomenon is that our text data set may be too small. Recall, in this experiment we mainly use glosses from different on-line dictionaries. These texts are of high precision, but are usually short and may not reflect all the semantic connections between actions. Another reason is more intrinsic. For some action pairs like walking and running, clapping and waving, they are “neighbors” in both the video and text spaces. On the other hand, the complex “jumping jacks” is quite different from all the other actions, as a result the auxiliary data and text data set appear to help less in this case.

Cross-lingual sentiment analysis

As labeling is expensive, it has long been a topic of interest in the NLP community to reuse previously labeled data for new applications. One recent effort is to transfer knowledge gained from data gathered in one language to another language. One obvious solution is to use the more and more popular automatic machine translation services (Banea et al. 2008). However, while machine translation has been the subject of a great deal of development in recent years, many of the recent gains in performance manifest as syntactically as opposed to semantically correct sentences. This causes problems especially when we are dealing with semantically oriented tasks such as the sentiment analysis problem we examine here. It is therefore still necessary to consider heterogeneous transfer learning when automatic translators are available.

Data set In this section we consider the binary sentiment classification problem. Given a review about a certain product, we want to predict whether it is positive or negative. As in our activity recognition setting, we will only have labeled data in the source feature space. But we will have some unlabeled data in the target space. We use the same training and test set as Wan (2009).

Model For this problem we have both X and Z as binary n-gram features. The joint model and gradient update is the same as we outline in the previous experiment. The remaining problem is that we do not have a natural auxiliary problem set in this case. We therefor need to construct these auxiliary problems in a way that we can automatically obtain labels for them.

We follow the work of Blitzer, Dredze, and Pereira (2007) by selecting a set of features (n-grams), which are highly related to the sentiment classification problem. The selection is performed by measuring the mutual information between features and the sentiment label variables y_{target} on the labeled data set in the source feature space (labeled English reviews). The auxiliary tasks are binary decision problems for these selected features. That is: whether a certain selected feature occurs in a text.

To label the data set with these auxiliary problems, we need the co-occurrence data $(y, \{x_i\})$ which is between a selected feature and a certain document in the target space. This can be obtained either by translating the whole document or simply referring to on-line dictionaries as we only need feature level co-occurrence. Note that our approach has less requirement on these co-occurrence statistics compared to Dai et al. (2008) as in their translator setting, all the feature pairs need to be considered in theory. While in our case, only a small number of co-occurrence data is used(only between normal features and pivot features).

To better estimate the performance of SCL and RBMs, we fix the Chinese unlabeled data set and sample different English labeled data sets from Blitzer, Dredze, and Pereira. We then calculate the average overall accuracy and the standard deviation of 20 experiment runs.

Result We use the results from Wan (2009) and Wei and Pal (2010) for comparison as we are using exactly the same training and test data set. The result of Wan was obtained by first translating the source label data into the target language, then using co-training scheme, which is a multi-view learning algorithm, on classifiers trained in both languages and let the two classifiers teach each other for the unlabeled examples. We use “co-train” for this result in the table. While the result from Wei and Pal was obtained by adapting the SCL algorithm, which is also a multi-task algorithm. We use “SCL” for the results. Finally, our results for the RBM approach is the average of 20 runs. By comparing our result with the two previous approaches, we can see how well our method is compared to the multi-view approach and other multi-task style algorithm.

From the figure, we can see that although our method doesn’t use as much information as Wan from the machine translation, we obtain a comparable and slightly better result.

	CoTrain	**SCL**	**RBM**
Pre(P)	76.8%	77.2%	77.0%
Rec(P)	90.5%	91.4%	90.9%
F1(P)	83.1%	83,8%	83.4%
Pre(N)	87.9%	93.1%	90.0%
Rec(N)	71.7%	75.2%	74.0 %
F1(N)	79.0%	83.3%	81.2 %
Overall	81.3%	83.5%	82.9%

Table 3: Results on cross-lingual sentiment analysis, Pre stands for precision, Rec for recall and F1 stands for F1 score; (P) means results on positive reviews, (N) on negative reviews

Method	μ	δ
SCL	83.2 %	0.13 %
RBM	83.0 %	0.10 %

Table 4: Average accuracy and Deviation, μ denotes the average overall accuracy, and δ stands for the standard deviation

The performance of RBMs in the specific experiment setting is slightly worse than the SCL approach. To more accurately estimate the performance of SCL and RBMs, we run 20 experiments on different English unlabeled data. And the result from Table 4 suggest that RBMs and SCL are statistically close to each other in the cross-langauge adaptation task.

Conclusion

In this paper, we have examined a problem setting we call heterogeneous transfer learning, a special case of transfer learning where the transfer is performed on two non-overlapping feature spaces. Different from the earlier work, our approach relaxes the requirement of co-occurrence data in training. Instead, we build the connection between the two feature spaces by considering their relations to a set of related tasks. We also show that simple RBMs can be useful in transfer learning. We study two problems as experiments: text-aided activity recognition and cross-lingual sentiment classification. In both cases, the RBM method outperforms the previous multi-view learning approaches. The performance on the cross-lingual experiment is slightly worse than the result obtained in the SCL. Yet repetitive experiments show the results of RBMs and SCL are statistically similar.

In this paper, we perform our experiments using a simple line search with the exact gradient. While more aggressive gradient approximation schemes such as contrastive divergence are commonly used for RBMs, for heterogeneous transfer learning with discriminative RBMs it may be worthwhile to explore if other optimization approaches might further boost performance.

In transfer learning, there is the possibility that the two tasks are unrelated, in this case, the knowledge obtained may not help, and even hurt. For heterogeneous transfer learning, this risk of negative transfer is even higher, as the different feature space makes it harder to decide the similarity between the source and the target task. We believe future work exploring this issue could be fruitful. We also wish to perform experiments on other complex real world data.

References

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*.

Banea, C.; Mihalcea, R.; Wiebe, J.; and Hassan, S. 2008. Multilingual subjectivity analysis using machine translation. In *EMNLP*, 127–135.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the ACL.*

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the ACL.*

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*.

Dai, W.; Chen, Y.; Xue, G.-R.; Yang, Q.; and Yu, Y. 2008. Translated learning: Transfer learning across different feature spaces. In *Proceedings of NIPS*.

Larochelle, H., and Bengio, Y. 2008. Classification using discriminative restricted boltzmann machines. In *Proceedings of ICML.*

Prettenhofer, P., and Stein, B. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the ACL.*

Tran, D., and Sorokin, A. 2008. Human activity recognition with metric learning. In *Proceedings of ECCV*.

Wan, X. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL-AFNLP*, 235–243. Suntec, Singapore: Association for Computational Linguistics.

Wang, X.; Pal, C.; and McCallum, A. 2007. Generalized component analysis for text with heterogeneous attributes. In *KDD*.

Wei, B., and Pal, C. 2010. Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the ACL.*

Welling, M.; Rosen-Zvi, M.; and Hinton, G. E. 2004. Exponential family harmoniums with an application to information retrieval. In *Proceedings of NIPS*.

Yang, J.; Liu, Y.; Xing, E. P.; and Hauptmann, A. G. 2007. Harmonium models for semantic video representation and classification. In *Proceedings of SDM*.

Yang, Q.; Chen, Y.; rong Xue, G.; Dai, W.; and Yu, Y. 2008. Heterogeneous transfer learning for image clustering via the social web. In *Proceedings of ACL.*