A Domain Adaptation Regularization for Denoising Autoencoders

Stéphane Clinchant, Gabriela Csurka and Boris Chidlovskii
Xerox Research Centre Europe
6 chemin Maupertuis, Meylan, France
Firstname.Lastname@xrce.xerox.com

Abstract

Finding domain invariant features is critical for successful domain adaptation and transfer learning. However, in the case of unsupervised adaptation, there is a significant risk of overfitting on source training data. Recently, a regularization for domain adaptation was proposed for deep models by (Ganin and Lempitsky, 2015). We build on their work by suggesting a more appropriate regularization for denoising autoencoders. Our model remains unsupervised and can be computed in a closed form. On standard text classification adaptation tasks, our approach yields the state of the art results, with an important reduction of the learning cost.

1 Introduction

Domain Adaptation problem arises each time when we need to leverage labeled data in one or more related *source* domains, to learn a classifier for unseen data in a *target* domain. It has been studied for more than a decade, with applications in statistical machine translation, opinion mining, part of speech tagging, named entity recognition and document ranking (Daumé and Marcu, 2006; Pan and Yang, 2010; Zhou and Chang, 2014).

The idea of finding domain invariant features underpins numerous works in domain adaptation. A shared representation eases prediction tasks, and theoretical analyses uphold such hypotheses (Ben-David et al., 2007). For instance, (Daumé and Marcu, 2006; Daumé, 2009) have shown that replicating features in three main subspaces (source, common and target) yields improved accuracy as the classifier can subsequently pick the most relevant common features. With the pivoting technique (Blitzer et al., 2006; Pan et al., 2010), the bag of words features are projected on a subspace that captures the relations between some central *pivot* features and the remaining words. Similarly, there are several extensions of topic models and matrix factorization techniques where the latent factors are shared by source and target collections (Chen and Liu, 2014; Chen et al., 2013).

More recently, deep learning has been proposed as a generic solution to domain adaptation and transfer learning problems by demonstrating their ability to learn invariant features. On one hand, unsupervised models such as *denoising autoencoders* (Glorot et al., 2011) or models built on word embeddings (Bollegala et al., 2015) are shown to be effective for domain adaptation. On the other hand, supervised deep models (Long et al., 2015) can be designed to select an appropriate feature space for classification. Adaptation to a new domain can also be performed by fine tuning the neural network on the target task (Chopra et al., 2013). While such solutions perform relatively well, the refinement may require a significant amount of new labeled data. Recent work by (Ganin and Lempitsky, 2015) has proposed a better strategy; they proposed to regularize intermediate layers with a domain prediction task, i.e. deciding whether an object comes from the source or target domain.

This paper proposes to combine the domain prediction regularization idea of (Ganin and Lempitsky, 2015) with the denoising autoencoders. More precisely, we build on *stacked Marginalized Denoising Autoencoders* (sMDA) (Chen et al., 2012), which can be learned efficiently with a closed form solution. We show that such domain adaptation regularization keeps the benefits of the sMDA and yields results competitive to the state of the art results of (Ganin and Lempitsky, 2015).

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 26–31,
Berlin, Germany, August 7-12, 2016.

2 Target Regularized MDA

Stacked Denoising Autoencoders (sDA) (Vincent et al., 2008) are multi-layer neural networks trained to reconstruct input data from partial random corruption. The random corruption, called blank-out noise or *dropout*, consists in randomly setting to zero some input nodes with probability p; it has been shown to act as a regularizer (Wager et al., 2013). The sDA is composed of a set of stacked one-layer linear denoising autoencoder components, which consider a set of N input documents (represented by d-dimensional features $\mathbf{x}_n$) to be corrupted M times by random feature dropout and then reconstructed with a linear mapping $\mathbf{W} \in \mathbb{R}^{d \times d}$ by minimizing the squared reconstruction loss:

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^{N} \sum_{m=1}^{M} ||\mathbf{x}_n - \tilde{\mathbf{x}}_{nm}\mathbf{W}||^2. \qquad (1)$$

As explicit corruption comes at a high computational cost, (Chen et al., 2012) propose to *marginalize* the loss (1) by considering the limiting case when $M \to \infty$ and reducing *de facto* the learning cost. The main advantage of this method is a closed form solution for $\mathbf{W}$, which depends only on the uncorrupted inputs ($\mathbf{x}_n$) and the dropout probability. Several Marginalized Denoising Autoencoders (MDA) can be then stacked together to create a deep architecture where the representations of the $(l-1)^{th}$ layer serves as inputs to the l^{th} layer[1].

In the case of domain adaptation, the idea is to apply MDA (or sMDA) to the union of unlabeled source $\mathbf{X}^s$ and target $\mathbf{X}^t$ examples. Then, a standard learning algorithm such as SVM or Logistic Regression is trained on the labeled source data using the new feature representations ($\mathbf{x}_n^s\mathbf{W}$) which captures better the correlation between the source and target data.

In Figure 1, we illustrate the effect of the MDA; it shows the relation between the word log document frequency (x-axes) and the *expansion mass* defined as the total mass of words transformed into word i by MDA and represented by $\sum_j W_{ji}$. We can see that the mapping $\mathbf{W}$ learned by MDA is heavily influenced by frequent words. In fact, MDA behaves similarly to document expansion on text documents: it adds new words with a very small frequency and sometimes words with a small negative weight. As the figure shows, MDA promotes common words (despite the use of tf-idf weighting scheme) that are frequent both in source and target domains and hence aims to be domain invariant.

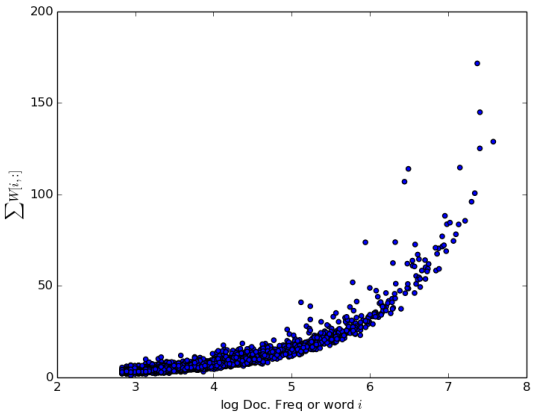

Figure 1: Relation between log document frequency and expansion mass. One dot represents one word.

This is in line with the work of (Ganin et al., 2015). To strengthen the invariance effect, they suggested a deep neural architecture which embeds a domain prediction task in intermediate layers, in order to capture domain invariant features. In this paper we go a step further and refine this argument by claiming that *we want to be domain invariant* but also to be *as close as possible to the target domain distribution*. We want to match the target feature distribution because it is where the classification takes place.

We therefore propose a regularization for the denoising autoencoders, in particular for MDA, with the aim to make *source data resemble the target data* and hence to ease the adaptation.

We describe here the case of two domains, but it can be easily generalized to multiple domains. Let $\mathbf{D}$ be the vector of size N indicating for each document its domain, *e.g.* taking values of '-1' for source and '$+1$' for target examples. Let $\mathbf{c}$ be a linear classifier represented as a d dimensional vector trained to distinguish between source and target, *e.g.* a ridge classifier that minimizes the loss $\mathcal{R}(\mathbf{c}, \alpha) = ||\mathbf{D} - \mathbf{X}\mathbf{c}^\top||^2 + \alpha||\mathbf{c}||^2$.

We guide the mapping $\mathbf{W}$ in such a way that the denoised data points $\mathbf{x}\mathbf{W}$ go towards the target side, *i.e.* $\mathbf{x}\mathbf{W}\mathbf{c}^\top = 1$ for both source and target

[1] Between layers, in general, a non linear function such as $tanh$ or $ReLU$ is applied.

samples. Hence, we can extend each term of the loss (1) as follows:

$$||\mathbf{x}_n - \tilde{\mathbf{x}}_{nm}\mathbf{W}||^2 + \lambda||\mathbf{1} - \tilde{\mathbf{x}}_{nm}\mathbf{W}\mathbf{c}^\top||^2. \quad (2)$$

The first term here represents the reconstruction loss of the original input, like in MDA. In the second term, $\tilde{\mathbf{x}}_{mn}\mathbf{W}\mathbf{c}^\top$ is the domain classifier prediction for the denoised objects forced to be close to 1, the target domain indicator, and $\lambda > 0$.

Let $\bar{\mathbf{X}}$ be the concatenation of M replicated version of the original data $\mathbf{X}$, and $\tilde{\mathbf{X}}$ be the matrix representation of the M corrupted versions. Taking into account the domain prediction term, the loss can be written as:

$$\mathcal{L}_{\mathcal{R}}(\mathbf{W}) = ||\bar{\mathbf{X}} - \tilde{\mathbf{X}}\mathbf{W}||^2 + \lambda||\bar{\mathbf{R}} - \tilde{\mathbf{X}}\mathbf{W}\mathbf{c}^\top||^2, \quad (3)$$

where $\mathbf{R}$ is a vector of size N, indicating a desired regularization objective, and $\bar{\mathbf{R}}$ its M-replicated version. Loss (3) represents a generic form to capture three different ideas:

- If $\mathbf{R} = \mathbf{1}$, the model incites the reconstructed features moving towards target specific features.
- If $\mathbf{R} = -\mathbf{D}$, the model aims to promote domain invariant features as in (Ganin et al., 2015).
- If $\mathbf{R} = [\mathbf{0}; \mathbf{1}]$, where $\mathbf{0}$ values are used for source data, the model penalizes the source specific features.

Learning the mapping W. (Chen et al., 2012) observed that the random corruption from equation (1) could be *marginalized out* from the reconstruction loss, yielding a unique and optimal solution. Furthermore, the mapping $\mathbf{W}$ can be expressed in closed form as $\mathbf{W} = \mathbf{P}\mathbf{Q}^{-1}$, with:

$$\begin{aligned} \mathbf{Q}_{ij} &= \begin{cases} \mathbf{S}_{ij}q_iq_j, & \text{if } i \neq j, \\ \mathbf{S}_{ij}q_i, & \text{if } i = j, \end{cases} \\ \mathbf{P}_{ij} &= \mathbf{S}_{ij}q_j, \end{aligned} \quad (4)$$

where[2] $q = [1-p, \ldots, 1-p] \in \mathbb{R}^d$, p is the dropout probability, and $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ is the covariance matrix of the uncorrupted data $\mathbf{X}$.

The domain regularization term in (3) is quadratic in $\mathbf{W}$, the random corruption can still be *marginalized out* and the expectations obtained in closed form. Indeed, the mapping $\mathbf{W}$ which minimizes the expectation of $\frac{1}{M}\mathcal{L}_{\mathcal{R}}(\mathbf{W})$ is the solution of the following linear system[3]:

$$(\mathbf{P} + \lambda(1-p)\mathbf{X}^\top\mathbf{R}\mathbf{c}^\top)(I + \lambda\mathbf{c}\mathbf{c}^\top)^{-1} = \mathbf{Q}\mathbf{W}. \quad (5)$$

In (5), parameter λ controls the effect of the proposed target regularization in the MDA and the regularization on $\mathbf{c}$ is controlled by parameter α. This approach preserves the good properties of MDA, *i.e.* the model is unsupervised and can be computed in closed form. In addition, we can easily stack several layers together and add non-linearities between layers.

3 Experiments

We conduct unsupervised domain adaptation experiments on two standard collections: the Amazon reviews (Blitzer et al., 2011) and the 20Newsgroup (Pan and Yang, 2010) datasets.

From the Amazon dataset we consider the four most used domains: *dvd (D), books (B), electronics (E) and kitchen (K)*, and adopt the settings of (Ganin et al., 2015) with the 5000 most frequent common features selected for each adaptation task and a tf-idf weighting. We then use the Logistic Regression (LR) to classify the reviews.

Our previous experiments with MDA revealed that the MDA noise probability p needs to be set to high values (*e.g.* 0.9). A possible explanation is that document representations are already sparse and adding low noise has no effect on the features already equal to zero. Figure 2 shows the average accuracy for the twelve Amazon tasks, when we vary the noise probability p.

In addition, we observed that a *single layer* with a $tanh$ activation function is sufficient to achieve top performance; stacking several layers and/or concatenating the outputs with the original features yields no improvement but increases the computational time.

The dropout probability p is fixed to 0.9 in all experiments, for both the MDA baseline and our model; we test the performance with a single layer and a $tanh$ activation function. Stacking several layers is left for future experiments. Parameters α and λ are tuned on a grid of values[4] by cross validation on the source data. In other words, we

[2] In contrast to (Chen et al., 2012), we do not add a bias feature so that the domain and MDA have the same dimensionality. Experiments shown no impact on the performance.

[3] The derivation is not included due to space limitation.

[4] $\alpha \in [.1, 1, 50, 100, 150, 200, 300]$, $\lambda \in [.01, .1, 1, 10]$.

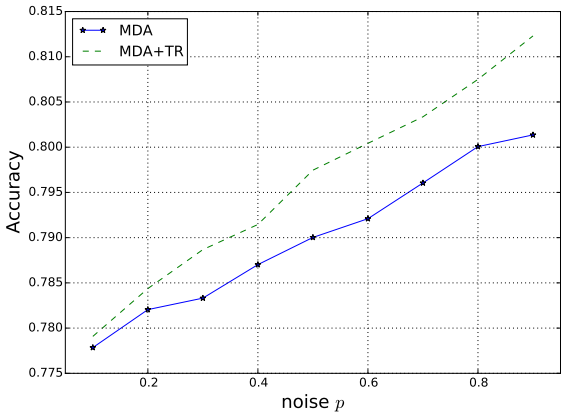

Figure 2: Impact of the noise parameter p on the average accuracy for the 12 Amazon adaptation tasks. Both MDA and its extension with the regularization (MDA+TR) perform better with a high dropout-out noise. Here MDA+TR is run with fixed parameters $\alpha = 100$ and $\lambda = 1$.

select the LR parameters and the parameters α, λ by cross validating the classification results using only the "reconstructed" source data; for estimating $\mathbf{W}$ we used the source with an unlabeled target set (excluded at test time). This corresponds to the setting used in (Ganin et al., 2015), with the difference that they use SVM and *reverse cross-validation*[5].

Table 3 shows the results for twelve adaptation tasks on the Amazon review dataset for the four following methods. Columns 1 and 2 show the LR classification results on the target set for the single layer MDA and the proposed target regularized MDA (MDA+TR). Column 3 reports the SVM result on the target from (Ganin et al., 2015). They used a 5 layers sMDA where the 5 outputs are concatenated with input to generate 30,000 features, on which the SVM is then trained and tested (G-sMDA). Finally, column 4 shows the current state of the art results obtained with Domain-Adversarial Training of Neural Networks (DA_NN) instead of SVM (Ganin et al., 2015).

Despite a single layer and LR trained on the source only, the MDA baseline (80.15% on average) is very close to the G-sMDA results obtained with 5 layer sMDA and 6 times larger feature set (80.18%). Furthermore, adding the target regularization allows to significantly outperform in many

[5]It consists in using self training on the target validation set and calibrating parameters on a validation set from the source labeled data.

S	T	MDA	MDA+TR	G-sMDA	DA_NN
D	B	81.1	81.4	82.6	**82.5**
D	K	84.1	**85.3**	84.2	84.9
D	E	76.0	**81.1**	73.9	80.9
B	D	82.7	81.7	**83.0**	82.9
B	K	79.8	81.8	82.1	**84.3**
B	E	75.9	79.3	76.6	**80.4**
K	D	78.5	**79.0**	78.8	**78.9**
K	B	**77.0**	77.0	76.9	71.8
K	E	87.2	**87.4**	86.1	85.6
E	D	**78.5**	78.3	77.0	78.1
E	B	73.3	75.1	76.2	**77.4**
E	K	87.7	**88.2**	84.7	**88.1**
Avg		80.15	81.27	80.18	**81.32**

Table 1: Accuracies of MDA, MDA+TR, G-sMDA and DA_NN on the Amazon review dataset. Underline indicates improvement over the baseline MDA, bold indicates the highest value.

cases the baseline and the state of the art DA_NN. We note that our method has a much lower cost, as it uses *the closed form solution* for the reconstruction and a *simple LR on the reconstructed source data*, instead of domain-adversarial training of deep neural networks.

We also look at the difference between the previously introduced expansion mass for the MDA and MDA+TR. In the adaptation task from *dvd* (D) to *electronics* (E), the words for which the mass changed the most are the following[6]: *worked, to_use, speakers, i_have, work, mouse, bought, cable, works, quality, unit, ipod, price, _number_, sound, card, phone, use, product, my*. These words are mostly target specific and the results confirm that they get promoted by the new model.

Our model favors features which are more likely to appear in target examples, while DA_NN seeks domain invariant features. Despite this difference, the two approaches achieve similar results. It is surprising, and we argue that eventually both approaches penalize *source specific features*. To test this hypothesis, we use MDA with $\mathbf{R} = [\mathbf{0}; \mathbf{1}]$ (case 3) that penalizes source specific features and we obtain again similar performances.

Finally, we test our approach on the 20Newsgroup adaptation tasks described in (Pan and Yang, 2010). We first filter out rare words and keep at most 10,000 features. Then, we apply both MDA and MDA+TR as above. Table 3 shows results for ten adaptation tasks. As we can see, in all cases the target regularization (MDA+TR) helps improve the classification accuracy.

[6]In ascending order of the differences.

Task	MDA	MDA+TR
comp vs sci	73.69	73.38
sci vs comp	69.39	69.92
rec vs talk	72.54	85.10
talk vs rec	72.30	76.22
rec vs sci	77.25	82.70
sci vs rec	79.95	80.00
sci vs talk	78.94	79.26
talk vs sci	77.17	77.91
comp vs rec	89.84	89.66
rec vs comp	89.92	90.29
Avg	78.1	80.40

Table 2: Accuracies of MDA and MDA+TR on 20Newsgroup adaptation tasks.

4 Conclusion

This paper proposes a domain adaptation regularization for denoising autoencoders, in particular for marginalized ones. One limitation of our model is the linearity assumption for the domain classifier, but for textual data, linear classifiers are the state of the art technique. As new words and expressions become more frequent in a new domain, the idea of using the dropout regularization that forces the reconstruction of initial objects to resemble target domain objects is rewarding. The main advantage of the new model is in the closed form solution. It is also unsupervised, as it does not require labeled target examples and yields performance results comparable with the current state of the art.

References

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems, NIPS Conference Proceedings, Vancouver, British Columbia, Canada, December 4-7, 2006.*, volume 19.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP, 22-23 July 2006, Sydney, Australia*.

John Blitzer, Sham Kakade, and Dean P. Foster. 2011. Domain adaptation with coupled subspaces. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS, Fort Lauderdale, USA, April 11-13, 2011*.

Danushka Bollegala, Takanori Maehara, and Ken-ichi Kawarabayashi. 2015. Unsupervised cross-domain word representation learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL, July 26-31, 2015, Beijing, China*, volume 1.

Zhiyuan Chen and Bing Liu. 2014. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31st International Conference on Machine Learning, ICML Bejing, 21-16 June 2014.*

M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha. 2012. Marginalized denoising autoencoders for domain adaptation. *ICML*, arXiv:1206.4683.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Leveraging multi-domain prior knowledge in topic models. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 2071–2077. AAAI Press.

S. Chopra, S. Balakrishnan, and R. Gopalan. 2013. DLID: Deep learning for domain adaptation by interpolating between domains. In *Proceedings of the 30th International Conference on Machine Learning, ICML, Atlanta, USA, 16-21 June 2013*.

H. Daumé and D. Marcu. 2006. Domain adaptation for statistical classifiers. *JAIR*, 26:101–126.

H. Daumé. 2009. Frustratingly easy domain adaptation. *CoRR*, arXiv:0907.1815.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML, Lille, France, 6-11 July 2015*, pages 1180–1189.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2015. Domain-adversarial training of neural networks. *CoRR*, abs/1505.07818.

X. Glorot, A. Bordes, and Y. Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning, ICML, Bellevue, Washington, USA, June 28-July 2, 2011*.

M. Long, Y. Cao, , J. Wang, and M. Jordan. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*.

S. J. Pan and Q. Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW, New York, NY, USA. ACM.

P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25nd International Conference on Machine Learning, ICML, Helsinki, Finland on July 5-9, 2008.*

Stefan Wager, Sida I. Wang, and Percy Liang. 2013. Dropout training as adaptive regularization. In 26, editor, *Advances in Neural Information Processing Systems, NIPS Conference Proceedings, Lake Tahoe, Nevada, United States, December 5-8, 2013.*

Mianwei Zhou and Kevin C. Chang. 2014. Unifying learning to rank and domain adaptation: Enabling cross-task document scoring. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 781–790, New York, NY, USA. ACM.