A General Regularization Framework for Domain Adaptation

Wei Lu[1] and **Hai Leong Chieu**[2] and **Jonathan Löfgren**[3]
[1]Singapore University of Technology and Design
[2]DSO National Laboratories
[3]Uppsala University
luwei@sutd.edu.sg, chaileon@dso.org.sg, lofgren021@gmail.com

Abstract

We propose a domain adaptation framework, and formally prove that it generalizes the feature augmentation technique in (Daumé III, 2007) and the multi-task regularization framework in (Evgeniou and Pontil, 2004). We show that our framework is strictly more general than these approaches and allows practitioners to tune hyper-parameters to encourage transfer between close domains and avoid negative transfer between distant ones.

1 Introduction

Domain adaptation (DA) is an important problem that has received substantial attention in natural language processing (Blitzer et al., 2006; Daumé III, 2007; Finkel and Manning, 2009; Daumé III et al., 2010). In this paper, we propose a novel regularization framework which allows DA practitioners to tune hyper-parameters to encourage transfer between close domains, and avoid negative transfer (Rosenstein et al., 2005) between distant ones. In our framework, model parameters in multiple domains are learned jointly and constrained to remain close to one another. In the transfer learning taxonomy (Pan and Yang, 2010), our framework falls under the parameter-transfer category for multi-task inductive learning. We show that our framework generalizes the frustratingly easy domain adaptation (FEDA) in Daumé III (2007), Finkel and Manning (2009), and the regularised multi-task learning of Evgeniou and Pontil (2004). At the same time, it provides us with hyper-parameters to control the amount of transfer between domains.

2 Domain Adaptation Framework

Given labeled data from N domains, $\mathbf{D}^1, \ldots, \mathbf{D}^N$, traditional machine learning maximizes the following objective function for each domain $\mathbf{D}^i$:

$$\mathcal{O}(\mathbf{D}^i; \mathbf{w}_i) = \mathcal{L}_i(\mathbf{D}^i; \mathbf{w}_i) - \lambda_i ||\mathbf{w}_i||^2, \quad (1)$$

and we maximize $\mathcal{L}_i$ by tuning the parameter vector $\mathbf{w}_i$. For example, $\mathcal{L}_i$ can be the log-likelihood or the negative hinge loss. The term $\lambda_i ||\mathbf{w}_i||^2$ is the L_2-regularization term where λ_i is a positive scalar. In our framework, we propose to maximize

$$\sum_{i=1}^{N} \mathcal{L}_i(\mathbf{D}^i; \mathbf{w}_i) - \sum_{i=1}^{N} \eta_{0,i} ||\mathbf{w}_i||^2 - \sum\nolimits_{1 \le j < k \le N} \eta_{j,k} ||\mathbf{w}_j - \mathbf{w}_k||^2, \quad (2)$$

where $\eta_{j,k}$ are parameters controlling the transfer between domains. In the next sections, we show how our framework generalizes existing works.

2.1 Frustratingly Easy DA

The FEDA approach was introduced by Daumé III (2007) and later formalized by Finkel and Manning (2009) within a hierarchical Bayesian DA framework. While simple, the approach has often been shown to be effective. In this section, we show that our framework generalizes the FEDA approach.

The FEDA approach defines a new augmented feature space by duplicating each feature in $\mathbf{D}^i$ to a "general" domain. Therefore each parameter in $\mathbf{w}_i$ has a corresponding parameter in $\mathbf{w}_0$, and:

$$\mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}_i, \mathbf{w}_0) = \mathcal{L}_i(\mathbf{D}^i; \mathbf{w}_i + \mathbf{w}_0) \quad (3)$$

This directly leads to the following remark:

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 950–954,
Austin, Texas, November 1-5, 2016.

Remark For all i, for any $\mathbf{w}_i, \mathbf{w}_0, \mathbf{d} \in \mathbf{R}^m$:

$$\mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}_i + \mathbf{d}, \mathbf{w}_0 - \mathbf{d}) = \mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}_i, \mathbf{w}_0)$$

The complete objective function involving N ($N \geq 2$) domains is defined as follows:

$$\mathcal{O}'(\mathbf{D}; , \mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_N) = \sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}_i, \mathbf{w}_0) - \sum_{i=0}^{N} \lambda_i ||\mathbf{w}_i||^2$$

We first prove the following relation:

Lemma 2.1 *Assume*

$$(\mathbf{w}_0^*, ..., \mathbf{w}_N^*) = \arg\max_{\mathbf{w}_1, ..., \mathbf{w}_N, \mathbf{w}_0} \left[\sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}_i, \mathbf{w}_0) - \left(\lambda_0 ||\mathbf{w}_0||^2 + \sum_{i=1}^{N} \lambda_i ||\mathbf{w}_i||^2 \right) \right],$$

where $\lambda_0, \lambda_1, \ldots, \lambda_N > 0$*, then:*

$$\lambda_0 \mathbf{w}_0^* = \sum_{i=1}^{N} \lambda_i \mathbf{w}_i^* \tag{4}$$

Proof Let's introduce the vector $\mathbf{d}$ as follows:

$$\mathbf{d} = \frac{1}{\sum_{i=0}^{N} \lambda_i} \left(\lambda_0 \mathbf{w}_0^* - \sum_{i=1}^{N} \lambda_i \mathbf{w}_i^* \right) \tag{5}$$

Denote $(\mathbf{w}'_0, \ldots, \mathbf{w}'_N)$ such that $\forall\ 0 \leq i \leq N$,

$$\mathbf{w}'_i = \mathbf{w}_i^* + \mathbf{d}, \text{ and } \mathbf{w}'_0 = \mathbf{w}_0^* - \mathbf{d}.$$

Based on the remark, $\mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}'_i, \mathbf{w}'_0) = \mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}_i^*, \mathbf{w}_0^*)$. Let $\Delta = \mathcal{O}'(\mathbf{D}; \mathbf{w}'_0, \ldots, \mathbf{w}'_N) - \mathcal{O}'(\mathbf{D}; \mathbf{w}_0^*, \ldots, \mathbf{w}_N^*)$. Since $(\mathbf{w}_0^*, \ldots, \mathbf{w}_N^*)$ is optimal, $\Delta \leq 0$. Moreover,

$$\begin{aligned}
\Delta &= \sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}'_i, \mathbf{w}'_0) - \sum_{i=0}^{N} \lambda_i ||\mathbf{w}'_i||^2 \\
&\quad - \sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i; \mathbf{w}_i^*, \mathbf{w}_0^*) + \sum_{i=0}^{N} \lambda_i ||\mathbf{w}_i^*||^2 \\
&= \lambda_0 ||\mathbf{w}_0^*||^2 - \lambda_0 ||\mathbf{w}_0^* - \mathbf{d}||^2 \\
&\quad + \sum_{i=1}^{N} \lambda_i ||\mathbf{w}_i^*||^2 - \sum_{i=1}^{N} \lambda_i ||\mathbf{w}_i^* + \mathbf{d}||^2 \\
&= -\left(\sum_{i=0}^{N} \lambda_i \right) ||\mathbf{d}||^2 + \\
&\quad 2\mathbf{d} \cdot \left(\lambda_0 \mathbf{w}_0^* - \sum_{i=1}^{N} \lambda_i \mathbf{w}_i^* \right) \\
&= -\left(\sum_{i=0}^{N} \lambda_i \right) ||\mathbf{d}||^2 + 2\mathbf{d} \cdot \left(\sum_{i=0}^{N} \lambda_i \right) \mathbf{d} \\
&= \left(\sum_{i=0}^{N} \lambda_i \right) ||\mathbf{d}||^2 \geq 0
\end{aligned}$$

Hence, $\Delta = 0$ implying $||\mathbf{d}|| = 0$ and so $\mathbf{d} = \mathbf{0}$. From the definition of $\mathbf{d}$, Equation 4 holds. ∎

Next we state the following lemma (see supplementary material for the proof).

Lemma 2.2 *For any vectors* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N \in \mathbf{R}^m$*, any scalars* $\lambda_0, \lambda_1, \ldots, \lambda_N \in \mathbf{R}^+$*, let* $\mathbf{v}_0 = (\sum_{i=1}^{N} \lambda_i \mathbf{v}_i)/\lambda_0$*, then the following always holds:*

$$\lambda_0 ||\mathbf{v}_0||^2 + \sum_{i=1}^{N} \lambda_i ||\mathbf{v}_i||^2 = \sum_{i=1}^{N} \eta_{0,i} ||\mathbf{v}_i + \mathbf{v}_0||^2 + \sum_{1 \leq j < k \leq N} \eta_{j,k} ||\mathbf{v}_j - \mathbf{v}_k||^2,$$

$$\text{where} \quad \eta_{i,j} = \frac{\lambda_i \lambda_j}{\sum_{l=0}^{N} \lambda_l},\ \forall\ 0 \leq i < j \leq N.$$

Now we state and prove the following theorem, which shows our framework generalizes FEDA.

Theorem 2.3 *For* $\lambda_0, \lambda_1, \ldots, \lambda_N \in \mathbf{R}^+$*, define*

$$\forall 0 \leq i < j \leq N, \quad \eta_{i,j} = \frac{\lambda_i \lambda_j}{\sum_{l=0}^{N} \lambda_l},$$

the following holds:

$$\max_{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_N,\mathbf{w}_0} \left[\sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i;\mathbf{w}_i,\mathbf{w}_0) - \left(\lambda_0||\mathbf{w}_0||^2 + \sum_{i=1}^{N}\lambda_i||\mathbf{w}_i||^2\right)\right]$$
$$= \max_{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_N} \left[\sum_{i=1}^{N} \mathcal{L}_i(\mathbf{D}^i;\mathbf{w}_i) - \left(\sum_{i=1}^{N}\eta_{0,i}||\mathbf{w}_i||^2 + \sum_{1\le j<k\le N}\eta_{j,k}||\mathbf{w}_j-\mathbf{w}_k||^2\right)\right]$$

Proof Let $(\mathbf{w}_0^*,\ldots,\mathbf{w}_N^*)$ be a solution to the first optimization problem. We have:

$$\mathbf{LHS} = \sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i;\mathbf{w}_i^*,\mathbf{w}_0^*) - \left(\lambda_0||\mathbf{w}_0^*||^2 + \sum_{i=1}^{N}\lambda_i||\mathbf{w}_i^*||^2\right) \quad (6)$$

Lemma 2.1 gives $\mathbf{w}_0^* = \left(\sum_{i=1}^{N}\lambda_i\mathbf{w}_i^*\right)/\lambda_0$. Introduce $\mathbf{w}'_i = \mathbf{w}_i^* + \mathbf{w}_0^*$. Using Lemma 2.2, we have:

$$\mathbf{LHS} = \sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i;\mathbf{w}_i^*,\mathbf{w}_0^*) - \left(\sum_{i=1}^{N}\eta_{0,i}||\mathbf{w}_i^*+\mathbf{w}_0^*||^2 + \sum_{1\le j<k\le N}\eta_{j,k}||\mathbf{w}_j^*-\mathbf{w}_k^*||^2\right)$$
$$= \sum_{i=1}^{N} \mathcal{L}_i(\mathbf{D}^i;\mathbf{w}'_i) - \left(\sum_{i=1}^{N}\eta_{0,i}||\mathbf{w}'_i||^2 + \sum_{1\le j<k\le N}\eta_{j,k}||\mathbf{w}'_j-\mathbf{w}'_k||^2\right) \le \mathbf{RHS}$$

Now, let $(\mathbf{w}_1^*,\mathbf{w}_2^*,\ldots,\mathbf{w}_N^*)$ be an optimal solution to the second problem. Given the relation between $\eta_{i,j}$ and $\lambda_0,\lambda_1,\ldots,\lambda_N$, let $\mathbf{w}'_0 = \left(\sum_{i=1}^{N}\lambda_i\mathbf{w}_i^*\right)/\left(\sum_{l=0}^{N}\lambda_l\right)$, and $\mathbf{w}'_i = \mathbf{w}_i^* - \mathbf{w}'_0$. We show in the supplementary material that

$$\mathbf{w}'_0 = \frac{1}{\lambda_0}\left(\sum_{i=1}^{N}\lambda_i\mathbf{w}'_i\right) \quad (7)$$

Based on these and Lemma 2.2, we have:

$$\mathbf{RHS} = \sum_{i=1}^{N} \mathcal{L}_i(\mathbf{D}^i;\mathbf{w}_i^*) - \left(\sum_{i=1}^{N}\eta_{0,i}||\mathbf{w}_i^*||^2 + \sum_{1\le j<k\le N}\eta_{j,k}||\mathbf{w}_j^*-\mathbf{w}_k^*||^2\right)$$
$$= \sum_{i=1}^{N} \mathcal{L}_i(\mathbf{D}^i;\mathbf{w}'_i+\mathbf{w}'_0) - \left(\sum_{i=1}^{N}\eta_{0,i}||\mathbf{w}'_i+\mathbf{w}'_0||^2 + \sum_{1\le j<k\le N}\eta_{j,k}||\mathbf{w}'_j-\mathbf{w}'_k||^2\right)$$
$$= \sum_{i=1}^{N} \mathcal{L}'_i(\mathbf{D}^i;\mathbf{w}'_i,\mathbf{w}'_0) - \left(\lambda_0||\mathbf{w}'_0||^2 + \sum_{i=1}^{N}\lambda_i||\mathbf{w}'_i||^2\right) \le \mathbf{LHS}$$

Therefore we must have $\mathbf{LHS} = \mathbf{RHS}$. ∎

This formally shows that FEDA is equivalent to solving the objective function given in Equation 2. In this new optimization problem, if we drop the terms involving $\eta_{j,k}$ for $j \neq 0$, we have:

$$\sum_{i=1}^{N}\left(\mathcal{L}_i(\mathbf{D}^i;\mathbf{w}_i) - \eta_{0,i}||\mathbf{w}_i||^2\right) \quad (8)$$

This is learning without domain adaptation. The additional regularization terms allow us keep the parameters from different domains close to one other. In the special case with two domains, if we use the same λ for all regularization terms, we have the following corollary:

Corollary 2.4 *For any $\lambda > 0$:*

$$\max_{\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_0}\left[\mathcal{L}'_1(\mathbf{D}^1;\mathbf{w}_1,\mathbf{w}_0) + \mathcal{L}'_2(\mathbf{D}^2;\mathbf{w}_2,\mathbf{w}_0) - \lambda\left(||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_0||^2\right)\right]$$
$$= \max_{\mathbf{w}_1,\mathbf{w}_2}\left[\mathcal{L}_1(\mathbf{D}^1;\mathbf{w}_1) + \mathcal{L}_2(\mathbf{D}^2;\mathbf{w}_2) - \frac{1}{3}\lambda\left(||\mathbf{w}_1||^2 + ||\mathbf{w}_2||^2 + ||\mathbf{w}_1-\mathbf{w}_2||^2\right)\right]$$

Hence, the FEDA feature augmentation technique indirectly introduces a regularization term that pushes the source and target parameters as close

as possible. This is related to the technique of Chelba and Acero (2006) where they regularize the model parameters for the target domain using the term $\lambda||\mathbf{w} - \mathbf{w}^s||$, where $\mathbf{w}^s$ is the parameter vector learned from the source domain. The difference here is, in their work the parameters for the source domain are learned first and then fixed. The relation between their work and the feature augmentation technique was also briefly discussed in the paper of Daumé III (2007). We formally showed a precise relation here in this paper.

2.2 Regularized Multi-task Learning

Evgeniou and Pontil (2004) proposed multi-task regularized learning using support vector machines (SVM). They decomposed the model weight vector as a sum of domain-specific vectors and a general vector, in much the same way as FEDA[1]. Hence, both Lemma 2.1 and Theorem 2.3 of this paper apply, and our framework also generalizes multi-task regularized learning.

3 Experimental Results

In this section we apply our framework to both structured and un-structured tasks. For structured prediction, we use the named-entity recognition (NER) ACE-2005 dataset with 7 classes and 6 domains. We apply the linear chain CRF (Lafferty et al., 2001), and show results using standard and softmax-margin CRF (SM-CRF) (Gimpel and Smith, 2010), with features consisting of word shape features, neighboring words, previous prediction and prefixes/suffixes. The second task is sentiment classification on the Amazon review data set (Blitzer et al., 2007) from 4 domains, labeled positive or negative. We apply logistic regression (LR) and SVM using unigram and bigram features. All the models used in this section are implemented on top of a common framework, which was also used to implement various structured prediction models previously (Lu, 2015; Lu and Roth, 2015; Muis and Lu, 2016). For each task we compare:

TGT Trained only on the specific domain data,
ALL Trained on the data from all domains,

[1]They proved in Lemma 2.1 in their paper a similar relationship to Equation 4, but their proof assumes a SVM framework, and that λ_1=λ_2=. . . =λ_N.

Model	Dom.	TGT	ALL	AUG	RF
CRF	bc	71.85	75.56	75.30	**76.48**
	bn	72.06	75.02	**75.17**	75.15
	cts	85.49	85.98	86.44	**86.70**
	nw	72.55	76.52	76.27	**76.61**
	un	67.09	72.99	72.90	**73.12**
	wl	64.38	69.66	69.46	**69.90**
	avg	72.24	75.96	75.92	**76.33**
SM-CRF	bc	72.33	75.54	75.04	**76.50**
	bn	72.18	74.86	75.10	**75.44**
	cts	85.68	85.96	86.15	**86.89**
	nw	72.70	76.19	75.92	**76.50**
	un	66.83	**72.94**	72.91	72.93
	wl	64.57	69.90	69.76	**70.30**
	avg	72.38	75.90	75.81	**76.43**

Table 1: F-score on the ACE NER task. The domains are broadcast conversations (bc), broadcast news (bn), conversational telephone speech (cts), newswire (nw), usenet (un) and weblog (wl). The macro-average (avg) over the 6 domains is also shown in the table.

Model	Dom.	TGT	ALL	AUG	RF
LR	book	75.83	79.33	79.00	**80.67**
	dvd	82.17	82.83	**83.83**	**83.83**
	elec.	84.67	84.67	**84.83**	**84.83**
	kit.	83.83	86.33	86.17	**87.33**
	avg	81.63	83.29	83.46	**84.17**
SVM	book	76.83	80.67	80.33	**81.00**
	dvd	83.17	83.17	82.50	**84.00**
	elec.	85.00	**86.50**	85.83	85.67
	kit.	86.33	85.83	**88.33**	87.83
	avg	82.83	84.04	84.25	**84.63**

Table 2: Accuracies on the sentiment classification task. The domains are books (book), dvds (dvd), electronics (elec.) and kitchen (kit.). The macro-average (avg) over the four domains are also shown in the table.

AUG The FEDA approach, and
RF Our proposed regularization framework.

We use a 40/30/30 train-development-test split and report the results on the test set. The regularization parameters were tuned on the development set over a logarithmic scale between 10^{-3} to 10^3. For our framework, we used random search to tune the parameters, since an exhaustive search is too expensive (21 parameters for 6 domains). We choose the within-domain $\eta_{0,i}$ to be close to those used for the *ALL* and *AUG* model, while choosing the other $\eta_{j,k}$ to be 1-2 orders of magnitude higher. A good model could quickly be found that generally beats the baselines on the development set and also generalizes well to the test set. We show the results for NER in Table 1 and the sentiment task in Table 2.

4 Discussion

Our proof did not require any assumption about $\mathcal{L}$, as long as L_2 regularization is used. This means our result is applicable to a variety of models such as SVM, LR, and CRF (where L_2 regularization is used for the latter two models). Theoretically, we have shown the equivalence of DA optimization problems. Empirically, for non-convex objectives, different approaches may arrive at different solutions. However, for convex loss functions, our objective (Equation 2) is also convex, and all approaches should share the same solution.

We have shown that we can map the FEDA optimization problem to our framework. The converse is false: for any problem in this family (with arbitrary choices of η), we can only solve it using FEDA if there are only 2 domains, or if all regularization hyper-parameters are equal. Some parameter configurations in this family are "unreachable" by the feature augmentation technique. This is because in Theorem 2.3, the values of η's are defined based on λ's and therefore possess certain properties. For example, they must at least satisfy such constraints as $\eta_{i,k}\eta_{k,j} = \eta_{i,l}\eta_{l,j}$ for any $i \leq k, l \leq j$. We have seen that some of those unreachable problems could give us better empirical results. Can we find an alternative simple adaptation method such that all problems in this family are "reachable"? This is a question that needs to be addressed in future research.

5 Conclusion

In this paper, we presented a framework for domain adaptation that generalizes several previous works (Daumé III, 2007; Finkel and Manning, 2009; Evgeniou and Pontil, 2004). Our approach allows practitioners to specify the amount of transfer between domains via regularization hyper-parameters. These parameters could be tuned based on intuition or using held-out data. In future work we could also seek to find methods that can automatically optimize these parameters. The supplementary material of this paper is available at `http://statnlp.org/research/ml/`.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments, and Zhanming Jie for his help on this work. The experiments of this work were done when Jonathan Löfgren was a visiting student at Singapore University of Technology and Design (SUTD). This work is supported by MOE Tier 1 grant SUTDT12015008.

References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of EMNLP*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*.

Ciprian Chelba and Alex Acero. 2006. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of 2010 Workshop on Domain Adaptation for Natural Language Processing*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of KDD*.

J. R. Finkel and C.D. Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of ACL*.

Kevin Gimpel and Noah A. Smith. 2010. Softmax-margin crfs: Training log-linear models with cost functions. In *Proceedings of HLT-NAACL*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of EMNLP*.

Wei Lu. 2015. Constrained semantic forests for improved discriminative semantic parsing. In *Proceedings of ACL/IJCNLP*.

Aldrian Obaja Muis and Wei Lu. 2016. Weak semi-markov crfs for noun phrase chunking in informal text. In *Proceedings of NAACL*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October.

Michael T. Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G. Dietterich. 2005. To transfer or not to transfer. In *In NIPS'05 Workshop, Inductive Transfer: 10 Years Later*.