Effective Multiclass Transfer For Hypothesis Transfer Learning

No Author Given

No Institute Given

Abstract. In this paper, we investigate the visual domain adaptation problem under the setting of *Hypothesis Transfer Learning* (HTL) where we can only access the source model instead of the data. We aim at two important issues: effectiveness of the transfer on small target training set and compatibility of the transfer model for real-world HTL problems. To solve these two issues, we proposed our method, Effective Multiclass Transfer Learning (EMTLe). We demonstrate that EMTLe, which uses the prediction of the source models as the transferable knowledge can exploit the knowledge of different types of source classifiers. We use the transfer parameter to weight the importance the prediction of each source model as the auxiliary bias and the output of the model trained from target model is adjusted with the auxiliary bias. Since the transfer parameter cannot be solved directly, we use the bi-level optimization to estimate the transfer parameter. Specifically, we show that our bi-level optimization problem is convex and we can effectively obtain the optimal transfer parameter with our novel objective function. Empirical results show that EMTLe can effectively exploit the knowledge and outperform other HTL baselines when target training set is small.

1 Introduction

Domain adaptation for image recognition tries to exploit the knowledge from a source domain with plentiful data to help learn a classifier for the target domain with a different distribution and little labeled training data. In domain adaptation, the source and target domains share the same label but their data are drawn from different distributions.

In domain adaptation, the knowledge of the source domain can be transferred by 3 different approaches: *instance transfer*, *model transfer* and *feature representation transfer* [13]. In this paper, we focus on the model transfer approach. Some recent works show that exploiting the knowledge from the source model can boost the performance of the target model effectively[11, 16]. Moreover, in some real applications, we can only obtain the source models and it is difficult to access their training data for different reasons such as the data credential. Recently, a framework called Hypothesis Transfer Learning (HTL) [10] has been proposed to handle this situation. HTL assumes only source models trained on the source domain can be utilized and there is no access to source data, nor any knowledge about the relatedness of the source and target distributions.

Previous research [2, 3] shows that without carefully measuring the distribution similarity between the source and target data, the source knowledge could not be exploited effectively or even hurt the learning process (called *negative transfer*)[13]. However, as we are not able to access the source data in an HTL setting, how to effectively and safely exploit the knowledge from the source model could be an important issue in HTL, especially when target data is relatively small (Effectiveness issue). Moreover, the source models from different domains can be trained with different kinds of classifiers. For example most models trained from ImageNet are deep convolutional neural networks while some models of the VOC recognition task could be SVMs or ensemble models. Therefore, a practical HTL algorithm should be compatible with different types of source classifiers (Compatibility issue). To the best of our knowledge, none of the previous work in HTL is able to solve these two issues at the same time.

In this paper, we propose our method, called Effective Multiclass Transfer Learning (EMTLe), that can solve these two issues simultaneously. Previous work [8] suggests that using the prediction of the source model as the transferable knowledge can greatly increase the compatibility of the transfer model for the HTL problem. To solve the compatibility issue, we introduce our strategy that uses the class prediction of the source model as the transferable knowledge to help the classification. Specifically, we use the weighted class probabilities produced by the source models to adjust the prediction from the target model. Here we call the weight of each source model *transfer parameter* which essentially controls the amount of knowledge transferred from the specific model. We argue that the transfer parameter is a hyperparameter of our model and cannot be solved directly.

To estimate the transfer parameter, we introduce bi-level optimization[14], which has been widely used for many different hyperparameter optimization problems recently. Specifically, on the low-level optimization problem, we use a least-square SVMs to train a model on the target data and on the high level, we introduce our novel multi-class hinge loss with ℓ_2 penalty that can better estimate the transfer parameter when training set is small. Moreover, we show that our bi-level optimization transfer parameter estimation problem is a strongly convex optimization problem and demonstrate that our method EMTLe can find the $O(\log(t)/t)$ optimal solution with t iterations.

We perform comprehensive experiments on 4 real-world datasets from two benchmark datasets (3 from Office and 1 from Caltech256). We show that EMTLe can effectively transfer the knowledge with different types of source models and outperforms the baseline methods under the HTL setting.

2 Related Work

As we focus on the model transfer approach under the HTL setting, in this section, we review some important methods using this approach. A model transfer approach assumes that the parameters of the model for the source task can be transferred to the target task. Two types of learning methods are generally used

for model knowledge transfer, generative probabilistic method and max margin method.

generative probabilistic method can predict the target domain by combining the source distribution to generate a posterior distribution. Li et al [7] used Bayesian transfer learning approach to learn the common prior for object recognition. Davis et al.[6] used an approach based on a form of second-order Markov logic to compensate for the domain shift. Wang et al.[17] proposed a method to change the marginal and conditional distributions smoothly to transfer the knowledge between tasks.

Alternatively, max margin methods try to use the hyperplane parameter to transfer the knowledge between source and target domains. Yang et al.[19] proposed Adaptive SVMs transferring parameters by incorporating the auxiliary classifier trained from the source domain. In addition to Yang's work, Ayatar et al.[1] proposed PMT-SVM that can determine the transfer regularizer automatically according to the target data. Tommasi et al.[16] proposed Multi-KT that can utilize the parameters from multiple source models for the target classes . Kuzborskij et al.[11] proposed a similar method to learn new categories by leveraging the known source models. Luo et al.[8] proposed MKTL and used feature augmentation method to leverage the source model.

Our work corresponds to the context above. In this paper, we propose EMTLe based on the model transfer approach. Similar to [8], we focus on how to exploit the knowledge from the prediction of the source models but with simpler and effective learning strategy.

3 Using the Source Knowlege as the Auxiliary Bias

In this section, we introduce our strategy in EMTLe that can exploit the knowledge from different types of source classifiers. In general, for each example in the target domain, we use its output class probabilities from the source models as the auxiliary bias term to adjust the final prediction of the target model.

Suppose we have to recognition a image from one of the N visual classes and there are N experts each of who can only provide the probability of this image for one certain class (binary source model). After we make our decision for one example (prediction from target model), the experts provide their own decisions as well (probabilities from the source models). Their decisions can provide extra information regarding this example as the auxiliary bias and adjust our final prediction. As each of the expert is a specialist in one class, we should weigh their decisions as well due to the bias of their predictions (see Figure 1).

Here, the weight of each source model reflects the relatedness between the source model and our target domain. The more related they are, the better decision the source model can make and the larger weight we should apply to it. Specifically, in this paper, we call the weight *transfer parameter*. Therefore, for any target data $D = \{x, y\}$ and the given source models $f' = \{f'_1, ..., f'_N\}$, our goal is to find the target model f:

$$f = \arg\min_{f \in \mathcal{F}} \ell\left(f + \beta f', D\right) \tag{1}$$

where $\beta = [\beta_1, ..., \beta_N]$ is the transfer parameter and $\ell(\cdot, \cdot)$ is the loss function to learn the target model. It is obvious that assigning proper transfer parameter to the source model can significantly improve the performance of our final prediction. However, the transfer parameter in Eq.(1) is a hyperparameter and we cannot solve it directly. Therefore, we introduce our bi-level optimization method for transfer parameter estimation in the next section.

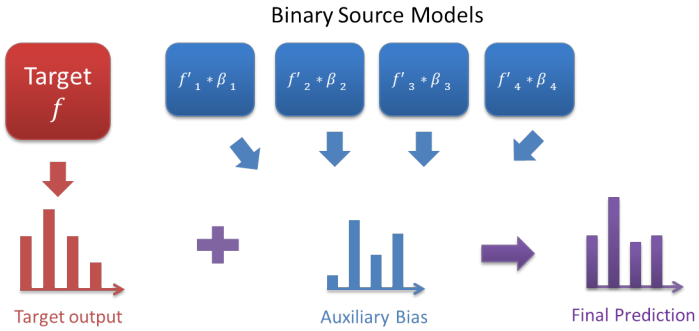


Fig. 1. Demonstration of using the source class probability as the auxiliary bias to adjust the output of the target model. f' is a group of binary classifiers $\{f'_1, ..., f'_4\}$ for each class and for each source model f'_n, we use the weight β_n to control the knowledge transferred from this model.

Unlike the previous work[1, 16, 18] which has to use the specific parameter of the source model as the source knowledge, our strategy is more compatible with different types of classifiers. Compared to [8] which uses sophisticated feature augmentation method to leverage the source model prediction, we provide a more straight-forward way to use the knowledge and have fewer hyperparameters to estimate as well. In addition, there are two advantages of our strategy: (1) It is an effective and easy way to align the knowledge from different types of source classifiers. (2) The auxiliary bias term is naturally normalized in the same dimension as the class probabilities are always in the interval $[0, 1]$. As EMTLe can select more types of source classifiers, this makes it more practical in a real HTL scenario.

From Eq. (1) we can see that, once we have determined the value of the transfer parameter β, we are able to find the target model f and solve the learning problem. In the next part, we will show how we can effective estimate the transfer parameter.

4 Bi-level Optimization for Transfer Parameter Estimation

As we discussed before, the transfer parameter in Eq. (1) is a hyperparameter that can't be solved directly. Here we use bi-level optimization (**BO**)[14], a popular method that is used for hyperparameter optimization to estimate the transfer parameter. In BO, the low-level optimization problem is to learn the target model and the high-level one is another cross-validation (CV) optimization problem for hyperparameter optimization corresponding to the model learned at low-level. Suppose we use K-fold CV on the high-level problem. For the i-th round CV, the target set D is split into training set D_i^{tr} and validation set D_i^{val}. The transfer parameter can be optimized with the following BO function:

$$\begin{aligned} \text{High level} \quad & \beta = \arg\min_{\beta} \sum_{i}^{K} \mathcal{L}(f^i(\beta), D_i^{val}) \\ \text{Low level} \quad & f^i(\beta) = \arg\min_{f \in \mathcal{F}} \ell\left(f + \beta f', D_i^{tr}\right) \end{aligned} \tag{2}$$

Here, $\ell(\cdot,\cdot)$ and $\mathcal{L}(\cdot,\cdot)$ are our low-level and high-level objective functions respectively. We can use any convex objective function in Eq.(2) for optimization (e.g. SVM objective function). In this paper, we use the leave-one-out cross-validation (**LOOCV**) in the high-level optimization problem. Previous research [10] suggests that LOOCV can increase the robustness of the estimated hyperparameter especially on the small dataset. In some previous works[12, 14], BO is a non-convex problem and can only obtain the approximate solution. However, we will show that problem (2) is strongly convex and we are able to obtain its optimal solution.

4.1 Low-level optimization problem using mean square loss

To better illustrate our learning scenario, we define our learning process as follows. Suppose we have N visual categories and can obtain N source binary classifiers $f' = \{f'_1, ..., f'_N\}$ from the source domain. We want to train a target function f consists of N binary classifiers $f = \{f_1, ..., f_N\}$ using the target training set D and the source models f'. Specifically, in our BO problem Eq. (2), for the low level optimization problem, we consider the scenario where we have to train N binary linear target model $f_i = w_i x + b_i$ so that for any $\{x_i, y_i\}_{i=1}^{l} \in D$, the adjusted result $f(x) + f'(x)\beta = y$. Let $D^{\setminus i} = D \setminus \{x_i, y_i\}$. Then, we use Least-square loss in the low-level objective function to optimize each target model f_n with any given transfer parameter β:

$$\begin{aligned} \text{Low-level:} \quad & f^{\setminus i}(\beta) : \min \sum_{n}^{N} \frac{1}{2}||w_n||^2 + \frac{C}{2} \sum_{j} \left(Y_{jn} - f_n(x_j) - \beta_n f'_n(x_j)\right)^2 \\ & \text{s.t.} \qquad f_n(x) = w_n x + b_n; \quad x_j \in D^{\setminus i} \end{aligned} \tag{3}$$

Here, Y is an encoded matrix of y using one-hot strategy where $Y_{in} = 1$ if $y_i = n$ and 0 otherwise.

The reason why we use the objective function (3) is that, it can provide an unbiased closed form Leave-one-out error estimation of each binary model f_n[4]. As a result, the high-level optimization problem becomes a convex optimization problem and we are able to estimate our transfer parameter easier.

Let $K(X, X)$ be the kernel matrix and

$$\psi = \left[K(X, X) + \frac{1}{C}\mathrm{I}\right] \tag{4}$$

Let ψ^{-1} is the inverse of matrix ψ and ψ_{ii}^{-1} is the ith diagonal element of ψ^{-1}. $\hat{Y}_{in}$, the LOO estimation of binary model $f_n^{\setminus i}$ for sample x_i, can be written as:

$$\hat{Y}_{in} = Y_{in} - \frac{\alpha_{in}}{\psi_{ii}^{-1}} \quad \text{for} \quad n = 1, ..., N \tag{5}$$

where the matrix $\boldsymbol{\alpha} = \{\alpha_{in} | i = 1, ...l; n = 1, ..., N\}$ can be calculated as:

$$\boldsymbol{\alpha} = \psi^{-1}Y - \psi^{-1}f'(X)\boldsymbol{\beta}^T \tag{6}$$

4.2 High-level optimization problem using multi-class hinge loss with ℓ_2 pentalty

For the high level optimization problem, we use multi-class hinge loss [5] with ℓ_2 penalty in our objective function.

$$\begin{aligned} \text{High-level:} \quad & \beta : \min \frac{\lambda}{2}\sum_{n}^{N} \|\beta_n\|^2 + \sum_{i,n}\left[1 - \varepsilon_{ny_i} + \hat{Y}_{in} - \hat{Y}_{iy_i} - \xi_i\right] \\ \text{s.t.} \quad & 1 - \varepsilon_{ny_i} + \hat{Y}_{in} - \hat{Y}_{iy_i} \leq \xi_i \end{aligned} \tag{7}$$

Here, $\varepsilon_{ny_i} = 1$ if $n = y_i$ otherwise 0. Compared to the previous work [11, 16] which uses the multi-class hinge loss without the ℓ_2 penalty, there are two main advantages for our high-level objective function: (1) When the training set is small, our LOOCV estimation could have a large variance. Similar to the penalty term in low-level problem (3), the ℓ_2 penalty here can reduce this variance and improve the generalization ability of the estimated transfer parameter. (2) It is clear that $\hat{Y}$ is a linear function w.r.t β. With the ℓ_2 penalty, optimization problem (7) become a strongly convex optimization problem w.r.t. the transfer parameter β. Therefore, we can obtain an $O(\log(t)/t)$ optimal solution with t iterations using Algorithm 1 (see proof in Theorem 1 in Appendix).

5 Experiment

In this section, we show empirical results of our algorithm for different transferring situations on two image benchmark datasets: Office and Caltech.

Algorithm 1 SMTLe

Input: $\lambda, \psi, Y, f', T,$
Output: $\beta = \{\beta^1, ..., \beta^n\}$
1: $\beta^0 = 1$,$\alpha' = \psi^{-1}Y, \alpha'' = \psi^{-1}f'$
2: **for** $t = 1$ to T **do**
3: $\hat{Y} \leftarrow Y - \left(\psi^{-1} \circ I\right)^{-1}(\alpha' - \alpha''\beta), \qquad \Delta_\beta = 0$
4: **for** $i = 1$ to l **do**
5: $\Delta_\beta \leftarrow \Delta_\beta + \lambda\beta$
6: $l_{ir} = \max(1 - \varepsilon_{y_i r} + \hat{Y}_{ir} - \hat{Y}_{iy_i})$
7: **if** $l_{ir} > 0$ **then**
8: $\Delta_\beta^{y_i} \leftarrow \Delta_\beta^{y_i} - \frac{\alpha''_{iy_i}}{\psi_{ii}^{-1}}$, $\Delta_\beta^{r} \leftarrow \Delta_\beta^{r} + \frac{\alpha''_{ir}}{\psi_{ii}^{-1}}$
9: **end if**
10: **end for**
11: $\beta^t \leftarrow \beta^{(t-1)} - \frac{\Delta_\beta}{\lambda \times t}$
12: **end for**

5.1 Dataset & Baseline methods

Office contains 31 classes from 3 subsets (Amazon,Dslr and Webcam) and Caltech contains 256 classes. We select 13 shared classes from two datasets[1]. The input features of all examples are extracted using AlextNet[9]. We compare our

Table 1. Statistics of the datasets and subsets

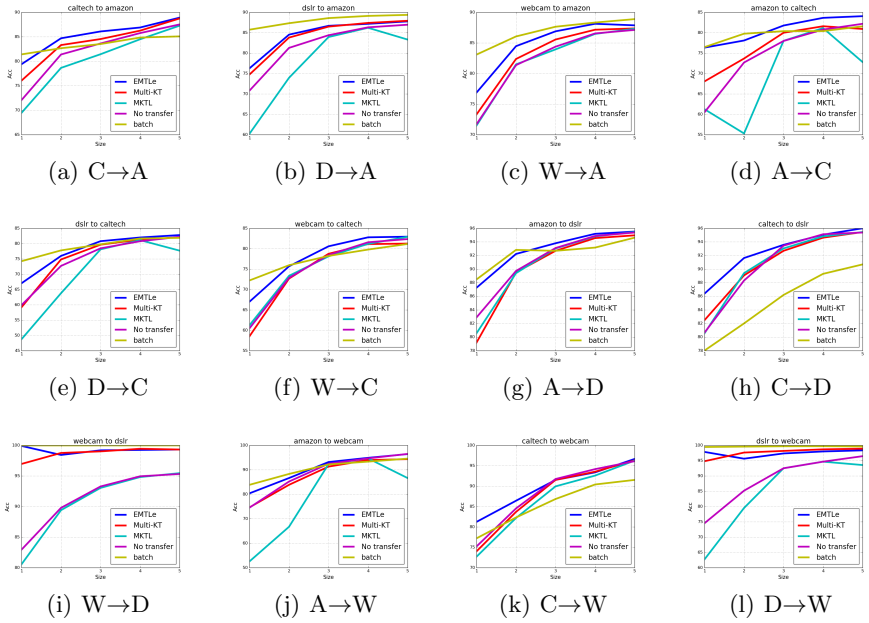
Dataset	Subsets	# classes	# examples	# features
Office	Amazon	13	1173	4096
	Dslr	13	224	4096
	Webcam	13	369	4096
Caltech256	Caltech	13	1582	4096

algorithm EMTLe with two kinds of baselines. The first one is the methods without leveraging any source knowledge (no transfer baselines), including two methods. **No transfer:** SVMs trained only on target data. Any transfer algorithm that performs worse than it suffers from negative transfer. **Batch:** We combined the source and target data, assuming that we have full access to all data, to train the SVMs. The result of the Batch method is expected to outperform other methods under the HTL setting as it can access the source data. The second kind of baseline consists of two previous transfer methods in HTL, **MKTL[8]** and **Multi-KT[16]**. Similar to EMTLe, both of them use the LOOCV method to estimate the relatedness of the source model and target domain, but they use their own convex objective function without the ℓ_2 penalty terms. We use linear kernel for all methods in all our experiments.

[1] 13 classes include: backpack, bike, helmet, bottle, calculator, headphone, keyboard, laptop, monitor, mouse, mug, phone and projector

5.2 Transfer from Single Source Domain

In this subsection, following the protocol in [8, 16] for fair comparison, we perform 12 groups of experiments under the setting of HTL. For each experiment, one of the 4 (sub)datasets is selected as the source, while another dataset is used as the target. We evaluate the the effectiveness of EMTLe when all source models are of the same type (linear SVMs). The size of each target dataset is varied from 1 to 5 to see how EMTLe and other baselines behave under the extremely small data setting. Experiment results are reported by averaging over 10 rounds and shown in Figure 2.



(a) C→A (b) D→A (c) W→A (d) A→C

(e) D→C (f) W→C (g) A→D (h) C→D

(i) W→D (j) A→W (k) C→W (l) D→W

Fig. 2. Recognition accuracy for HTL domain adaptation using single source single source classifier. 5 different sizes of target training sets are used in each group of experiments.

Observation & discussion: EMTLe can significantly outperform other baselines especially when the training size is small. As we have discussed above, when the training set is small, with the transfer parameter estimated by our ℓ_2 penalty in our high-level objective functions, EMTLe has a strong generalization ability and performs better on the test data. As the training size increases, the variance of training data decreases and the affect of the ℓ_2 penalty term become less significant. Therefore, EMTLe and the other two HTL baselines show similar performance. It is interesting to see that MKTL even fall into negative transfer even with 5 training examples per class in some experiments. We found

that, MKTL is more sensitive to variance of the training data its performance is not as stable as Multi-KT and EMTLe over the 10 rounds experiments. Because MKTL need to learn more hyperparameters than Multi-KT and EMTLe, even though the training size increases, it may not be able to obtain a good model. In some experiments, we can see that EMTLe can even outperform the Batch method which which can access more information and is expected to outperform the other methods under the setting of HTL.

5.3 Transfer from Multiple Source Domains

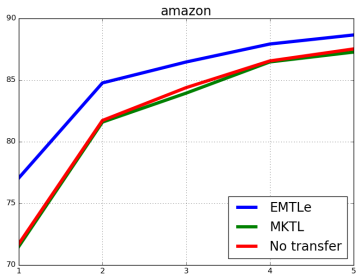
As we mentioned, EMTLe can exploit knowledge from different types of source classifiers which could greatly extend our selection of the source domain under the HTL setting. In this subsection we show that EMTLe can successfully transfer the knowledge from two source models of different types of source classifiers. Meanwhile, MKTL is used as our baseline which can also be compatible with different types of source classifiers.

In this experiment, we assume that there is no single source domain that can cover all classes in our target domain and we have to select source models from different source domains. Specifically, the 13 binary source models are selected from two different domains separately (6 from DSLR and 7 from Webcam) according to Table 2. Similar to our previous experiment configurations, we only use Caltech and Amazon as the target domain. We show the experiment results in Figure 3.

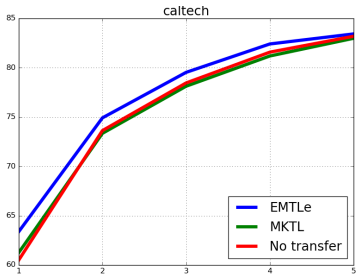
Table 2. The selected classes of the two source domains and the classifier type of the source model.

	class	classifier
DSRL	monitor,bike, helmet,calcu,headphone,projector	Logitic
Webcam	keyboard,mouse,phone,backpack,mug,bottle,laptop	SVMs

Observation & discussion: Under our multi-source scenario, it is more difficult leverage the knowledge from the source models as the models are trained from different domains. From the result we can see that, in this complex situation, EMTLe can still transfer the knowledge from the source models despite the type of the source classifiers while MKTL can hardly get any improvement. EMTLe uses a simple way to leverage the source models and BO can help us better estimate the transfer parameter. However, MKTL uses sophisticated feature augmentation to leverage the source models and has more hyperparameters to estimate. With a few training data, it is difficult for MKTL to measure the importance of each source model and exploit the knowledge from the models effectively.



(a) Dslr+Webcam → Amazon



(b) Dslr+Webcam → Caltech

Fig. 3. Recognition Accuracy for Multi-Model & Multi-Source experiment on two target datasets.

6 Conclusion

In this paper, we proposed a method, EMTLe that can effectively transfer the knowledge under the HTL setting. We focus on the effectiveness and compatibility issues for HTL problems. We proposed our auxiliary bias strategy to let our model exploit the knowledge from different types of source classifiers. The transfer parameter of EMTLe is estimated by bi-level optimization method using our novel high-level objective function which allows our model to better exploit the knowledge from source models. Experiment results demonstrate that EMTLe can effectively transfer the knowledge even though the size of training data is extremely small.

Appendix

Theorem 1. *Let $L(\beta)$ be a λ-strongly convex function and β^* be its optimal solution.Let $\beta_1, ..., \beta_{T+1}$ be a sequence such that $\beta_1 \in B$ and for $t > 1$, we have $\beta_{t+1} = \beta_t - \eta_t \Delta_t$, where Δ_t is the sub-gradient of $L(\beta_t)$ and $\eta_t = 1/(\lambda t)$. Assume we have $||\Delta_t|| \leq G$ for all t. Then we have:*

$$L(\beta_{T+1}) \leq L(\beta^*) + \frac{G^2(1 + \ln(T))}{2\lambda T} \tag{8}$$

Proof: As $L(\beta)$ is strongly convex and Δ_t is in its sub-gradient set at β_t, according to the definition of λ-strong convexity [15], the following inequality holds:

$$\langle \beta_t - \beta^*, \Delta_t \rangle \geq L(\beta_t) - L(\beta^*) + \frac{\lambda}{2}||\beta_t - \beta^*||^2 \tag{9}$$

For the term $\langle \beta_t - \beta^*, \Delta_y \rangle$, it can be written as:

$$\langle \beta_t - \beta^*, \Delta_t \rangle = \left\langle \beta_t - \frac{1}{2}\eta_t \Delta_t + \frac{1}{2}\eta_t \Delta_t - \beta^*, \Delta_t \right\rangle = \frac{1}{2}\langle \beta_{t+1} + \beta_t - 2\beta^*, \Delta_t \rangle + \frac{1}{2}\eta_t \Delta_t^2 \tag{10}$$

Then we have:

$$||\beta_t - \beta^*||^2 - ||\beta_{t+1} - \beta^*||^2 = \langle \beta_{t+1} + \beta_t - 2\beta^*, \eta_t \Delta_t \rangle \tag{11}$$

Using the assumption $||\Delta_t|| \leq G$, we can rearrange (9) and plug (10) and (11) into it, we have:

$$Diff_t = L(\beta_t) - L(\beta^*) \leq \frac{\lambda(t-1)}{2}||\beta_t - \beta^*||^2 - \frac{\lambda t}{2}||\beta_{t+1} - \beta^*||^2 + \frac{1}{2}\eta_t G^2 \tag{12}$$

Due to the strong convexity, for each pair of $L(\beta_t)$ and $L(\beta_{t+1})$ for $t = 1, ..., T$, according to (9), we have:

$$L(\beta_{t+1}) - L(\beta_t) \leq \langle \beta_{t+1} - \beta_t, \Delta_t \rangle - \frac{\lambda}{2}||\beta_{t+1} - \beta_t||^2 = -\eta_t \Delta_t^2 (1 - \frac{1}{2t}) \leq 0 \tag{13}$$

Therefore, we have the following sequence $L(\beta^*) \leq L(\beta_T) \leq L(\beta_{T-1}) \leq ... \leq L(\beta_1)$. For the sequence $Diff_t$ for $t = 1, ..., T$, we have:

$$\sum_{t=1}^{T} Diff_t = \sum_{t=1}^{T} L(\beta_t) - TL(\beta^*) \geq T\left[L(\beta_T) - L(\beta^*)\right] \tag{14}$$

Next, we show that

$$\begin{aligned} &\sum_{t=1}^{T} Diff_t = \sum_{t=1}^{T} \left\{ \frac{\lambda(t-1)}{2}||\beta_t - \beta^*||^2 - \frac{\lambda t}{2}||\beta_{t+1} - \beta^*||^2 + \frac{1}{2}\eta_t G^2 \right\} \\ &= -\frac{\lambda T}{2}||\beta_{T+1} - \beta^*||^2 + \frac{G^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{G^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t} \leq \frac{G^2}{2\lambda}(1 + \ln(T)) \end{aligned} \tag{15}$$

Combining (14) and rearranging the result, we have:

$$L(\beta_{T+1}) \leq L(\beta^*) + \frac{G^2(1 + \ln(T))}{2\lambda T}$$

References

1. Aytar, Y., Zisserman, A.: Tabula rasa: Model transfer for object category detection. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2252–2259. IEEE (2011)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning 79(1-2), 151–175 (2010)
3. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al.: Analysis of representations for domain adaptation. Advances in neural information processing systems 19, 137 (2007)

4. Cawley, G.C.: Leave-one-out cross-validation based model selection criteria for weighted ls-svms. In: Neural Networks, 2006. IJCNN'06. International Joint Conference on. pp. 1661–1668. IEEE (2006)
5. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. The Journal of Machine Learning Research 2, 265–292 (2002)
6. Davis, J., Domingos, P.: Deep transfer via second-order markov logic. In: Proceedings of the 26th annual international conference on machine learning. pp. 217–224. ACM (2009)
7. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(4), 594–611 (2006)
8. Jie, L., Tommasi, T., Caputo, B.: Multiclass transfer learning from unconstrained priors. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 1863–1870. IEEE (2011)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
10. Kuzborskij, I., Orabona, F.: Stability and hypothesis transfer learning. In: Proceedings of the 30th International Conference on Machine Learning. pp. 942–950 (2013)
11. Kuzborskij, I., Orabona, F., Caputo, B.: From n to n+ 1: Multiclass transfer incremental learning. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 3358–3365. IEEE (2013)
12. Maclaurin, D., Duvenaud, D., Adams, R.P.: Gradient-based hyperparameter optimization through reversible learning. In: Proceedings of the 32nd International Conference on Machine Learning (2015)
13. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
14. Pedregosa, F.: Hyperparameter optimization with approximate gradient. In: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. pp. 737–746 (2016)
15. Rockafellar, R.T.: Convex analysis. Princeton university press (2015)
16. Tommasi, T., Orabona, F., Caputo, B.: Learning categories from few examples with multi model knowledge transfer. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(5), 928–941 (2014)
17. Wang, X., Huang, T.K., Schneider, J.: Active transfer learning under model shift. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 1305–1313 (2014)
18. Yang, J., Yan, R., Hauptmann, A.G.: Adapting svm classifiers to data with shifted distributions. In: Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on. pp. 69–76. IEEE (2007)
19. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: Proceedings of the 15th international conference on Multimedia. pp. 188–197. ACM (2007)