Safety Multiclass Incremental Transfer Learning

Shuang Ao
Western University, Canada
sao@uwo.ca

Abstract

In transfer learning, Hypothesis Transfer Learning is proposed where we can only access to the models of the source domain rather than the data. In practice, in order to accept new category for HTL as the target task, we have to collect the data by ourselves. Without accessing the original data of the source domain, we have to evaluate the models for the models of original category as well. In order to solve this problem, we propose our method, called Safety Multiclass Incremental Transfer Learning (SMITLe). SMITLe can distinguish the utility of the source models effectively. We show that the transfer parameters learned from SMITLe can avoid negative transfer and SMITLe converges at a logarithmic rate. We design 3 sets of experiment that would happen in real learning scenario. Experimental results show that SMITLe can consistently achieve higher accuracy compared to previous methods.

1 Introduction

Transferring the knowledge between different image databases is a very popular topic in recent years due to the fast growing vision-based applications. There are many existing models to distinguish various objects. Exploiting the knowledge of these models properly can greatly help us to solve a specific recognition task. Recently, some works have been been proposed within a framework called Hypothesis Transfer Learning (HTL) [Tommasi *et al.*, 2014] [Kuzborskij *et al.*, 2013] [Jie *et al.*, 2011] [Kuzborskij and Orabona, 2013]. Unlike domain adaptation methods, HTL assumes only source models (called hypothesis) trained on source task can be utilized and there is no access to the data of source domain, nor any knowledge about the relatedness of the source and target distributions. Previous works of HTL focus on how to add a new category model to the existing N-category models, i.e. from a N-category (source) task to a $(N+1)$ (target) one. Previous work assumes that the hypotheses of the N categories are correct for the target task [Kuzborskij *et al.*, 2013]. However, for real application in HTL, since we are not able to access the source data, in order to add a new category model, we have to collect all the data of the $N+1$ categories by ourselves. Therefore, it is trivial to simply assume that the prior hypotheses is also correct for the data we collected, especially for image recognition tasks. For example, the source hypothesis is trained from to distinguish apple and orange using the images that are bright and have high contrast level, but the images of orange and apple collected by us could be low contrast and taken in a dark environment. Therefore, the source hypothesis may fail for the same target task. Therefore, we extend previous works by assuming that the hypotheses from source task may fail to classify the same category. When the source hypotheses fail, transferring from them could lead to negative transfer.

In transfer learning, when the data distribution of the source and target domain are different, transferring the knowledge between them could even degrade the performance of the classifier on target task, which is referred to as negative transfer. On the other hand, when the data distribution are similar for the two tasks, leveraging the knowledge from source task can improve the performance of the classifier for the target one(positive transfer). In our case, we are more likely to suffer from negative transfer due to the mismatched data distribution which could result from both the N original categories and the new added category (see Figure 1). How to avoid negative transfer is still an open question in transfer learning [Lu *et al.*, 2015]. Previous works suggest that, to better utilizing the hypotheses and reduce negative transfer, decision of the algorithm should be made by combining the prior hypotheses and empirical knowledge (from the specific target task) [Tommasi *et al.*, 2014] [Kuzborskij *et al.*, 2013] [Yang *et al.*, 2007] [Aytar and Zisserman, 2011].

In HTL, a number of empirical works have been attempted with Least Square Support Vector Machine (LS-SVM) [Kuzborskij and Orabona, 2013]. The framework of HTL with LS-SVM has two major phrases: (I) Building binary One-Versus-All SVMs with some biased regularization. (II) Estimating transfer parameters with some objective functions. Following these two phrases, we propose our method, called Safety Multiclass Incremental Transfer Learning (SMITLe), that can both avoid negative transfer and leverage correct hypothesis. In Phrase I, a regularization term adopted from Multi-KT [Tommasi *et al.*, 2014] is used to adapt the N original categories and the new category. As a result, the decision of each binary LS-SVM is the linear combi-

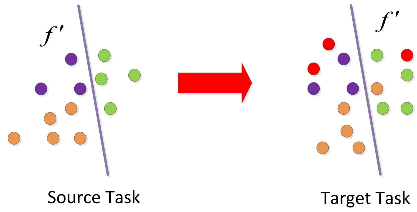

Figure 1: Negative transfer happens when we transfer prior hypothesis f' to target one. Points with different color represent different categories. The data distribution would change even for identical categories in different task. The new added category (red points) can also greatly affect the data distribution in target task.

nation of the prior hypotheses and empirical knowledge controlled by some transfer parameters. In Phrase II, to measure the transferability of each prior hypothesis, we estimate our transfer parameters using multi-class prediction error based on closed-form leave-one-out (LOO) error for model evaluation. Then we propose our novel objective function that can balance the weight between the prior hypotheses and empirical knowledge from target task. Experimental results show that SMITLe can achieve better accuracy than other baselines.

The main contributions of this paper include: (1) We propose a novel algorithm SMITLe within the HTL framework that can autonomously utilize the prior hypotheses to prevent negative transfer. (2) We also show that SMITLe can obtain the optimal solution at the rate of $O(\frac{\log(t)}{t})$ where t is training iteration.

The rest of this paper is organized as follow. In Section 2, we introduce the biased regularization terms of our problem for phrase 1 of HTL. Then, we propose a novel objective function for transfer parameter estimation, called SMITLe in Section 3. We show that the estimated transfer parameter can distinguish the utility of the prior hypothesis and avoid negative transfer autonomously. In Section 4, we show the performance comparison between SMITLe and other baselines on a variety of experiments on AwA and Caltech datasets in three different scenarios.

2 Biased Regularization

In this section, we focus on the Phrase I of HTL and introduce our biased regularization for binary LS-SVM for our problem. We use multi-source transfer strategy to generate our biased regularization. As a result, the decision of each binary LS-SVM is the linear combination of the knowledge from both target task and source hypotheses controlled by certain transfer parameters.

We define our task in the following way: assume that, for our $(N+1)$-category target task, $x \in \mathcal{X}$ and $y \in Y = \{1, 2, ..., N+1\}$ are the input vector and output for the learning task respectively. Meanwhile, we have a set of binary linear classifiers (hypotheses) $f'_n(x) = \phi(x)w'_n + b'_n$, for $n = 1, ..., N$ trained from an unknown distribution with One-Versus-All (OVA) strategy. Now we want to learn a set of classifiers $f_n(x) = \phi(x)w_n + b_n, n = 1, ..., N+1$ for our new task. The example x is assigned to the category j if $j \equiv \arg\max_{n=1,...,N+1} \{f_n(x)\}$. From previous works of HTL, the solution of the parameters (w_n, b_n), for each binary LS-SVM, can be found by solving the following optimization problem:

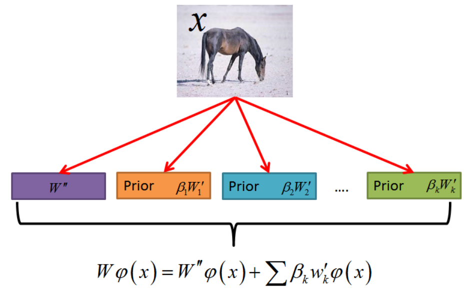

Figure 2: The final decision function value of a binary SVM can be get by combining the prior and empirical knowledge.

$$\mathbf{min} \quad R(w_n, W') + \frac{C}{2} \sum_{i}^{l} (Y_{i,n} - \phi(x_i)w_n - b_n)^2$$

Here, $W' = \{w'_1, w'_2, ..., w'_n\}$. $R(w_n, W')$ is the regularization term to guarantee good transfer performance and avoid overfitting. $\mathbf{Y}$ is a encoded label matrix so that $Y_{in} = 1$ if $y_i = n$ and -1 otherwise.

Now our task can be divided into two separate part: learning the the $(N+1)_{th}$ new category and N overlapped categories.

For the new added category, it is very difficult to identify the utility of the hypothesis of a single category in source task, therefore, we use multi-source transfer strategy, adopted from Multi-KT [Tommasi *et al.*, 2014], to leverage hypotheses from multiple sources. As a result, regularization term $R(w_{N+1}, W')$ can be written as:

$$R(w_{N+1}, W') = \frac{1}{2} \left\| w_{N+1} - \sum_{k=1}^{N} w'_k \beta_k \right\|^2 \tag{1}$$

We can interpret the biased regularization in the following way. Let $w_{N+1} = w''_{N+1} + \sum \beta_k w'_k$. Therefore, we have:

$$w_{N+1}\phi(x) = w''_{N+1}\phi(x) + \sum \beta_k w'_k \phi(x)$$

Here, we call β the transfer parameter. For any fixed value of β, regularizing w_n is equivalent to regularize w''_{N+1}, i.e. $(w_{N+1} - \sum \beta_k w'_k)$. The decision of each binary SVM model is made by combining the decision from target task $w''_{N+1}\phi(x)$ and source hypotheses $w'_k\phi(x)$ controlled by the transfer parameter. The amount of transfered knowledge has a positive correlation to the value of β.

However, for the original N categories, we already have their corresponding source category hypotheses and thus, their regularization term can be written as:

$$R(w_n, w'_n) = \frac{1}{2} \|w_n - \gamma_n w'{}_n\|^2 \tag{2}$$

As we can see that the regularization term (2) is a special case of (1) where only one β_k is none-zero.

Combining these two together, our multi-class incremental transfer problem can be solved by optimizing the following objective function:

$$\begin{aligned}\textbf{min} \quad & \frac{1}{2}\sum_{n=1}^{N}\|w_n-\gamma_n w'_n\|^2+\frac{1}{2}\left\|w_{N+1}-\sum_{k=1}^{N}w'_k\beta_k\right\|^2 \\ & +\frac{C}{2}\sum_{n=1}^{N+1}\sum_{i=1}^{l}e_{i,n}^2 \\ \textbf{s.t.} \quad & e_{i,n}=Y_{in}-\phi(x_i)w_n-b_n, \quad n\in\{1,...,N+1\}\end{aligned} \tag{3}$$

The optimal solution to Eq. (3) is:

$$\begin{aligned} w_n &= \gamma_n w'_n+\sum_{i}^{l}\alpha_{in}\phi(x_i), n=1,...,N \\ w_{N+1} &= \sum_{k}^{N}\beta_k w'_k+\sum_{i}^{l}\alpha_{i(N+1)}\phi(x_i)\end{aligned} \tag{4}$$

Here α_{ij} is the element (i,j) in $\boldsymbol{\alpha}$.

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

$$\psi=\begin{bmatrix} K(X,X)+\frac{1}{C}\mathrm{I} & \mathbf{1} \\ \mathbf{1^T} & 0\end{bmatrix} \tag{5}$$

$$\psi\begin{bmatrix}\boldsymbol{\alpha}' \\ \boldsymbol{b}'\end{bmatrix}=\begin{bmatrix} Y \\ 0\end{bmatrix}, \quad \psi\begin{bmatrix}\boldsymbol{\alpha}'' \\ \boldsymbol{b}''\end{bmatrix}=\begin{bmatrix} X(W')^T \\ 0\end{bmatrix} \tag{6}$$

We have:

$$\boldsymbol{\alpha}=\boldsymbol{\alpha}'-\begin{bmatrix}\boldsymbol{\alpha}''\boldsymbol{d_\gamma} & \boldsymbol{\alpha}''\boldsymbol{\beta^T}\end{bmatrix} \tag{7}$$

Here $\boldsymbol{d_\gamma}$ is a diagonal matrix with $\{\gamma_i\}_{i=1,...,N}$ in its main diagonal. From Eq. (7) we can see that, the solution of Eq. (3) is completed once $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are set.

3 SMITLe

In this section, we focus on the Phrase II of HTL, to estimate the transfer parameter in our task. We introduce an algorithm, called SMITLe, that can effectively estimate unbiased transfer parameter from small training set.

3.1 Multi-class Prediction Loss with LOO

From Phrase I, we can see that the amount of knowledge transfered is determined by the transfer parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Generally, we would like to reduce the amount of transfer from the prior hypotheses when they are incorrect. Meanwhile, for those correct ones, aggressively increasing the amount of transfer can boost the performance for the target problem. Once we fix the value of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, our task can be directly solved.

To evaluate different settings of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, we have to their cross-validation error iteratively. In this paper, we choose the Leave-One-Out (LOO) cross-validation error as the evaluation criterion. We choose it for the following two reasons: (1) It is proven that LOO error has low bias on small training data regime [Kuzborskij and Orabona, 2013]. (2) Moreover, it is exhausted to really perform cross-validations and compare the results for each setting of (γ,β). An important advantage of choosing LS-SVM over the other model is that we can obtain unbiased LOO error in closed form without real performing it.

The unbiased LOO estimation for sample x_i can be written as [Cawley, 2006]:

$$\hat{Y}_{i,n}=Y_{i,n}-\frac{\alpha_{in}}{\psi_{ii}^{-1}} \quad \text{for} \quad n=1,...,N+1 \tag{8}$$

Here ψ^{-1} is the inverse of matrix ψ and ψ_{ii}^{-1} is the ith diagonal element of ψ^{-1}.

Let us call ξ_i the multi-class prediction error for example x_i. ξ_i can be defined as [Crammer and Singer, 2002]:

$$\xi_i(\gamma,\beta)=\max_{n\in\{1,...,N+1\}}\left[1-\varepsilon_{ny_i}+\hat{Y}_{in}(\gamma,\beta)-\hat{Y}_{iy_i}(\gamma,\beta)\right] \tag{9}$$

Where $\varepsilon_{ny_i}=1$ if $n=y_i$ and 0 otherwise. The intuition behind this loss function is to enforce the distance between the true class and other classes to be at least 1.

Now, we already have an effective way to measure the performance of different settings of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ for our task. In the next part, we introduce how we optimize these parameters.

3.2 Loss Function of SMITLe

In this part, we propose a novel objective function according to our multi-class prediction loss function for transfer parameter estimation. We show that we can effectively obtain the optimal $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ that is resistant to negative transfer.

From (9) we can see that, different from the binary scenario where 0 is used as the hard threshold to distinguish the two classes, our multi-class loss only depends on the gap between the decision function value of the correct label ($\hat{Y}_{y_i}$) and the maximum among the decision function value of the other labels $\hat{Y}_{in}(n\neq y_i)$. To reduce ξ_i for a specific example x_i, we only have to increase the gap between $\hat{Y}_{in}(n\neq y_i)$ and $\hat{Y}_{iy_i}$.

As we mentioned before, the amount of knowledge transfered is positively correlated to the value of transfer parameter. When the prior hypotheses are correct, we have $w'_{y_i}\phi(x_i)>w'_n\phi(x_i)$. If $\xi_i>0$, increasing the transfer parameters can reduce the gap between $\hat{Y}_{y_i}$ and $\hat{Y}_{in}(n\neq y_i)$, leading to smaller ξ_i. When the prior hypotheses are incorrect and $\xi_i>0$, there exists a $j(j\neq y_i)$ such that $w'_{y_i}\phi(x_i)<w'_j\phi(x_i)$. Thus, reducing the transfer parameter can eventually reduce ξ_i.

Instead of optimize ξ_i directly, we add two extra regularization terms for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Then we define our objective function as:

$$\begin{aligned}\textbf{min} \quad & \frac{\lambda_1}{2}\sum_{n=1}^{N}\|\gamma_n\|^2+\frac{\lambda_2}{2}\sum_{n=1}^{N}\|\beta_n\|^2+\sum_{i=1}^{l}\xi_i \\ \textbf{s.t.} \quad & 1-\varepsilon_{ny_i}+\hat{Y}_{in}(\gamma,\beta)-\hat{Y}_{iy_i}(\gamma,\beta)\leq\xi_i; \\ & \lambda_1,\lambda_2\geq 0\end{aligned} \tag{10}$$

Here λ_1 and λ_2 are two regularization parameters to prevent negative transfer. By adding these two regularization terms, the objective function (10) turns to be strongly convex. Therefore, its strongly convex property guarantees that SMITLe can converge at the rate of $O(\frac{\log(t)}{t})$ (see proof in Appendix A).

From the objective function above we can see that, for certain λ_1 and λ_2, when the prior hypotheses are incorrect and harmful, decreasing $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ leads to smaller loss from both regularization and multi-class prediction error for target task. Moreover, we also prove that with optimal $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ from this objective function, SMITLe can actually avoid negative transfer (see Appendix B). On the other hand, if the prior hypotheses are incorrect, even though, increasing $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ leads to larger regularization loss, it also leads to smaller multi-class prediction error on the target problem. Therefore, the algorithm compromises between them.

3.3 Optimizing $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$

By adding a dual set of variables in objective function (10), one for each constraint in, we get the Lagrangian of the optimization problem:

$$
\begin{aligned}
&L(\gamma, \beta, \xi, \eta) = \frac{\lambda_1}{2}\sum_{n=1}^{N}\|\gamma_n\|^2 + \frac{\lambda_2}{2}\sum_{n=1}^{N}\|\beta_n\|^2 + \sum_{i=1}^{l}\xi_i \\
&+ \sum_{i,n}\eta_{i,n}\left[1 - \varepsilon_{ny_i} + \hat{Y}_{in}(\gamma,\beta) - \hat{Y}_{iy_i}(\gamma,\beta) - \xi_i\right] \\
&\textbf{s.t.} \quad \forall i, n \quad \eta_{i,n} \geq 0
\end{aligned} \tag{11}
$$

To obtain the optimal values for the problem above, we introduce our method using sub-gradient descent [Boyd and Vandenberghe, 2004] and summarize it in Algorithm. 1.

Algorithm 1 SMITLe optimization

Input: $\psi, \alpha', \alpha'', T, \psi,$
Output: $\gamma = \{\gamma^1, ..., \gamma^n\}, \beta$
$\beta^0 \leftarrow 0, \gamma^0 \leftarrow 1$
for $t = 1$ to T **do**
  $\hat{Y} \leftarrow Y - (\psi \circ I)^{-1}\left(\alpha' - \begin{bmatrix} \alpha'' d_\gamma & \alpha''\beta^T \end{bmatrix}\right)$
  $\Delta_\gamma = 0, \Delta_\beta = 0$
  for $i = 1$ to l **do**
    $\Delta_\gamma \leftarrow \Delta_\gamma + \lambda_1\gamma$, $\Delta_\beta \leftarrow \Delta_\beta + \lambda_2\beta$
    for $r = 1$ to $N + 1$ **do**
      $l_{ir} = 1 - \varepsilon_{y_i r} + \hat{Y}_{ir} - \hat{Y}_{iy_i}$
      if $l_{ir} > 0$ **then**
        if $y_i, r \in \{1, ..., N\}$ **then**
          $\Delta_\gamma^{y_i} \leftarrow \Delta_\gamma^{y_i} - \frac{\alpha''_{iy_i}}{\psi_{ii}^{-1}}$, $\Delta_\gamma^r \leftarrow \Delta_\gamma^r + \frac{\alpha''_{ir}}{\psi_{ii}^{-1}}$
        else if $y_i = N + 1$ **then**
          $\Delta_\beta \leftarrow \Delta_\beta - \frac{\alpha''_i}{\psi_{ii}^{-1}}$, $\Delta_\gamma^r \leftarrow \Delta_\gamma^r + \frac{\alpha''_{ir}}{\psi_{ii}^{-1}}$
        else
          $\Delta_\gamma^{y_i} \leftarrow \Delta_\gamma^{y_i} - \frac{\alpha''_{iy_i}}{\psi_{ii}^{-1}}$, $\Delta_\beta \leftarrow \Delta_\beta + \frac{\alpha''_i}{\psi_{ii}^{-1}}$
        end if
      end if
    end for
  end for
  $\beta^t \leftarrow \beta^{(t-1)} - \frac{\Delta_\beta}{l \times t}$, $\gamma^t \leftarrow \gamma^{(t-1)} - \frac{\Delta_\gamma}{l \times t}$
end for

4 Experiment

In this section, we show empirical results of our algorithm on different transferring situations on two datasets: AwA10[1] [Lampert *et al.*, 2009] and Caltech10[2] [Griffin *et al.*, 2007]. In real world applications, there are three situations in HTL. The first two extreme cases are all the hypotheses are correct/incorrect. The third one is the intermediate (mixed) case where only part of the hypotheses are correct. We design three sets of experiment, called positive, negative and mixed transfer experiment respectively, based on these 3 situations, comparing our algorithm with the baselines.

4.1 Dataset

Caltch10 is a subset of Caltech256. We select the following 10 categories: *bat, bear, dolphin, giraffe, gorilla, horse, leopard, raccoon, skunk, zebra*, containing 1387 images, as our dataset. AwA10 is a sub set of AwA dataset. Its source images is not publicly accessible and we can only access the six pre-extracted feature representations for each image. This property makes it natural as the unknown distribution source dataset to train the prior knowledge. We choose the identical 10 categories as those in Caltech10 from it, containing 6917 images.

[1]The features of AwA dataset is available from http://attributes.kyb.tuebingen.mpg.de/

[2]Images for Caltech is available from http://www.vision.caltech.edu/Image_Datasets/Caltech256/

4.2 Baselines and algorithmic setup

We compare our algorithm with two kinds of baselines. The first one is methods without leveraging any prior knowledge (no transfer baselines). The second consists of some methods with transfer techniques. Here are the no transfer baselines.

No transfer: LS-SVM 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 fully access to all data, to train the LS-SVM. The result of this baseline might be considered as the best performance achieved when the source and target tasks are related.

Source+1: This method only train a new binary LS-SVM for the new category. For the rest of the classes, we use the predictions of the classifiers trained from source data directly. The performance of this method can be an indicator whether the data of the source and target task are drawn from similar distribution.

We select the 3 HTL methods, MKTL [Jie *et al.*, 2011], MULTI-KT [Tommasi *et al.*, 2014] and MULTIpLE [Kuzborskij *et al.*, 2013], as our transfer baselines.

For all the experiments in this section, we adopt the same strategy as [Kuzborskij *et al.*, 2013] and [Tommasi *et al.*, 2014], using kernel averaging [Gehler and Nowozin, 2009] to compute the average of RBF kernels over the available features on RBF hyperparameter $\{2^{-5}, 2^{-4}, ..., 2^{8}\}$.

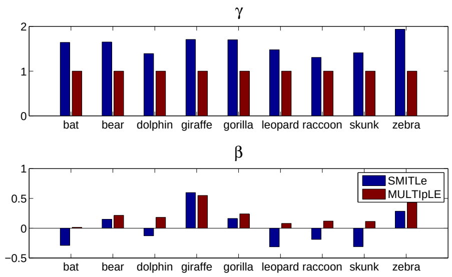

Figure 3: Experiment results for 10 classes, AwA. Horse is used as the new category. We can see that SMITLe tends to more aggressively exploit the related prior knowledge.

The penalty parameter C is tuned via cross-validation on $\{10^{-5}, 10^{-4}, ..., 10^{8}\}$ and the optimal value is reused for all the algorithms. Two transfer regularization parameters $\lambda 1$ and $\lambda 2$ are also set via cross-validation on $\{10^{-3}, 10^{-2}, ..., 10\}$ respectively.

4.3 Positive transfer: transferring from related sources

In the extreme case, where the knowledge of the source task is related to the target one, the data of these two tasks can be drawn from the same distribution. Thus we perform tow experiments under this setting on both AwA and Caltech. For each dataset, we split the data into two sets. One is treated as the source dataset to train the source model and another is treated as the target dataset for training and testing. We iteratively choose one category as the new category and run the experiment 10 times to get the average performance of each algorithm. The results of the two experiments are reported in Table 1 and Table 2. From the results we can see that, in Caltech experiment, our algorithm consistently outperforms all the baselines (even better than Batch method). In AwA dataset, Source+1 outperforms SMITLe when the training size is 5. As we increase the training size, the accuracy of SMITLe increases and outperforms Source+1.

To illustrate the detail performance of our algorithm, we select the experiment result on AwA dataset where horse is chosen as the new category for further explanation. In Figure 3 we provide values of γ and β compared with the parameters of the runner-up transfer algorithm MULTIpLE. We can see that for transfer knowledge between identical categories, MULTIpLE fixes the transfer parameter (γ) to be 1 while our method sets greater weights for related prior knowledge. By exploiting the positive prior knowledge more aggressively, SMITLe is able to leverage the prior knowledge and outperforms other methods. For the transfer parameter β we can see that MULTIpLE tends to keep β greater than 0 and SMITLe works more intuitively, setting positive weight for related categories(giraffe, zebra and bear etc.) and small or even negative weight for unrelated categories (bat, dolphin and skunk etc.).

Table 1: Average accuracy in percentage across all categories from Caltech to Caltech with different size of training set in target problem. 30 examples are randomly chosen from each class to train the source classifier and 30 examples from each class are chosen for test.

size per category	5	10	15	20
No transfer	27.33	31.53	35.73	38.47
Source+1	43.33	43.87	44.33	44.57
MKTL	38.89	43.27	45.72	47.44
MULTIKT	37.96	42.89	45.96	47.32
MULTIpLE	42.63	45.63	47.81	48.73
SMITLe	**43.53**	**46.45**	**48.25**	**49.15**
Batch	43.77	44.73	46.67	48.00

Table 2: Average accuracy in percentage across all categories from AwA to AwA with different size of training set in target problem. 50 examples are randomly chosen from each class to train the source classifier and 200 examples from each class are chosen for test.

size per category	5	10	15	20
No transfer	23.52	26.79	29.60	31.50
Source+1	**39.00**	**39.34**	39.62	39.74
MKTL	31.46	34.76	37.41	38.81
MULTIKT	29.86	32.86	35.22	36.33
MULTIpLE	37.80	38.81	39.80	40.47
SMITLe	37.83	39.31	**40.37**	**41.09**
Batch	39.62	40.18	40.67	41.44

4.4 Negative transfer: transferring from unrelated sources

In this section, we show how our method performs in transferring knowledge between two different datasets, from AwA dataset to Caltech dataset. Following the settings in previous experiment, the source models are trained from AwA dataset and transfered to Caltech dataset. We iteratively select one category as the new one, running multiple times to get the average results for all the algorithms. We show the average performance of each algorithm in Table 3. We can see that negative transfer does happen when transferring the knowledge from AwA to Caltech for all the algorithms except for ours. From the performance of Source+1, we can see that applying the source models directly leads to poor performance. We can conclude that even though these two datasets share some categories, the data distribution of the feature representation for the same category is not consistent.

Still we take the experiment where horse is considered as the new category to see the detail performance of each algorithm. From here we can see that, not surprisingly, the accuracy of SMITLe shows similar accuracy to the no transfer baseline, while other methods suffer from negative transfer and perform even worse than no transfer baseline. In Figure 4 we show the parameters learned for each classes in SMITLe in comparison with MULTIpLE. We can see that, when

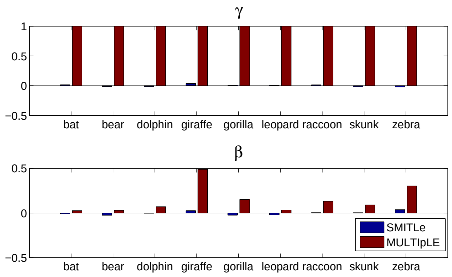

Figure 4: Experiment results for 10 classes, AwA. Horse is used as the new category. SMITLe can ignore unrelated prior knowledge.

the prior knowledge is unrelated, SMITLe resists utilizing the prior knowledge and therefore shows almost identical accuracy to the no transfer baseline.

Table 3: Average accuracy in percentage across all categories from AwA to Caltech. Examples in AwA are used to train prior models. Different number of training size is randomly selected from Caltech dataset.

	5	10	15	20	25
No transfer	**30.99**	33.97	35.95	37.78	38.27
Source+1	17.89	18.69	18.79	19.69	19.39
MKTL	25.19	30.14	32.53	34.30	35.83
MULTIKT	27.60	32.19	34.51	36.78	37.79
MULTIpLE	29.79	33.45	35.49	36.77	37.43
SMITLe	30.93	**34.13**	**36.09**	**38.01**	**38.46**

4.5 Transferring from mixed sources

In real applications, extreme situation is rare. For most multi-source transfer learning tasks, there should always be some related and useful sources as well as some unrelated ones. In this part, we show how SMITLe performs in the mixed sources.

From negative transfer experiments we see that the knowledge from AwA is unrelated to Caltech and vice versa. To generate mixed sources, we follow the settings in our positive transfer experiment, splitting the AwA dataset into two datasets, and replace the data of some categories in the source dataset with the data of Caltech. For example, if bat is considered as the new category and we have to replace 3 categories, we choose the data from 3 out of 9 categories (10 categories except for bat) in Caltech to replace the source data accordingly.

We show the performances across all categories of different algorithms in Table 4 and Table 5 where 3 and 4 categories in the source data are replaced by the data from Caltech respectively. From the tables we can see that in almost every case, SMITLe shows improved or equivalent performance than other baselines.

Table 4: Average accuracy in percentage across all categories from AwA to AwA&Caltech with different size of training set in target problem. Data of 3 classes in AwA is replaced by the data from Caltech in target problem.

	5	10	15	20	25
no transfer	23.99	26.24	29.02	30.05	31.18
source+1	25.70	26.30	26.57	26.69	26.97
MKTL	25.30	27.59	30.42	31.01	31.97
MultiKT	25.53	27.94	30.48	31.36	32.31
MULTIpLE	28.11	29.61	31.34	32.18	32.89
SMITLe	**28.75**	**30.48**	**32.30**	**33.06**	**33.71**

Table 5: Average accuracy in percentage across all categories from AwA to AwA&Caltech with different size of training set in target problem. Data of 4 classes in AwA is replaced by the data from Caltech in target problem.

	5	10	15	20	25
no transfer	24.02	26.25	29.06	30.07	31.20
source+1	23.23	23.80	24.03	24.21	24.47
MKTL	24.44	26.78	29.64	30.40	31.50
MultiKT	24.73	27.40	29.93	30.91	31.91
MULTIpLE	26.50	28.33	30.27	31.29	32.12
SMITLe	**27.20**	**29.33**	**31.40**	**32.31**	**33.11**

5 Conclusion

In this paper, we present a novel method called SMITLe that is able to transfer knowledge across different datasets and learn a new category. Inspired by previous work, SMITLe uses LS-SVM as the basic classification model and LOO for transfer parameter estimation. We demonstrate that SMITLe is able to converge at a logarithmic rate. We also prove that with the transfer parameters optimized by our novel objective function, SMITLe is able to avoid negative transfer which is a general issue for transfer learning. We carry out 3 sets of experiment that our algorithm would face in real world application. From the experimental results we can see SMITLe can consistently outperform other transfer baselines and achieve higher classification accuracy in different scenarios.

A Convergence Analysis

Let $\mu_1, ..., \mu_t$ be a sequence corresponding to $\mu_t = (\sqrt{\lambda_1}\gamma^t, \sqrt{\lambda_2}\beta^t)$. Problem (10) can be rewritten as:

$$J(\mu) = \frac{1}{2}\|\mu\|^2 + \sum_{i=1}^{l} \xi_i(\mu)$$

Let Δ_t be the sub-gradient for $J(\mu_t)$ and $\mu^* = (\sqrt{\lambda_1}\gamma^*, \sqrt{\lambda_2}\beta^*)$ be the optimal solution for it. Assume that $\|\Delta_t\| \leq G$. According to Lemma 1 in [Shalev-Shwartz *et al.*, 2011], we have:

$$J(\mu_t) - J(\mu^*) \leq \frac{G^2}{2t}(1 + \ln(t)) \qquad (12)$$

This means that SMITLe converges at the rate of $O(\frac{\log(t)}{t})$.

B Proof of avoiding negative transfer

Assume that $\bar{\xi}_i$ is the multi-class loss of example x_i without utilizing any prior knowledge, i.e. $\gamma = \beta = \mathbf{0}$. Let γ^*, β^* be the optimal solution for Eq. (11) and ξ_i^* be the multi-class loss with respective to example x_i. Then for every example $x_i \in \mathcal{X}$, we have:

$$\sum_i \xi_i \leq \sum_i \bar{\xi}_i$$

Proof. When $\gamma = \beta = \mathbf{0}$, from Eq. (9) we can get:

$$\bar{\xi}_i = \max_n \left[\varepsilon_{ny_i} - 1 + \frac{(\alpha'_{iy_i} - \alpha'_{in})}{\psi_{ii}^{-1}} \right]$$

For simplification, let $\delta_i = 1$ if $i = N + 1$ and 0 otherwise, and $\theta_{ij} = \alpha''_{ij} (1 - \delta_j) / \psi_{ii}^{-1}$. To find the minimum of the primal problem, we require:

$$\frac{\partial L}{\partial \xi_i} = 1 - \sum_n \eta_{in} = 0 \rightarrow \sum_n \eta_{in} = 1$$

$$\frac{\partial L}{\partial \gamma_n} = 0 \Rightarrow \gamma_n^* = \frac{1}{\lambda_1} \sum_i (\varepsilon_{ny_i} - \eta_{in}) \theta_{in} \tag{13}$$

$$\frac{\partial L}{\partial \beta_n} = 0 \Rightarrow \beta_n^* = \frac{1}{\lambda_2} \sum_{i,n} \frac{\eta_{in} \alpha''_{in}}{\psi_{ii}^{-1}} (\delta_{y_i} - \delta_n) \tag{14}$$

As the strong duality holds,the primal and dual objectives coincide. Plug Eq (13) and (14) into Eq. (11), we have:

$$\sum_{i,n} \eta_{in} \left[1 - \varepsilon_{ny_i} + \hat{Y}_{in} (\gamma^*, \beta^*) - \hat{Y}_{iy_i} (\gamma^*, \beta^*) - \xi_i^* \right] = 0$$

Expand the equation above, we have:

$$\begin{aligned} &\sum_{i,n} \eta_{in} \left[\varepsilon_{n,y_i} - 1 + \frac{(\alpha'_{iy_i} - \alpha'_{in})}{\psi_{ii}^{-1}} - \xi_i \right] \\ &\qquad = \lambda_1 \sum_r \|\gamma_r^*\|^2 + \lambda_2 \sum_r \|\beta_r^*\|^2 \geq 0 \end{aligned}$$

Rearranging the above, we obtain:

$$\sum_{i,n} \eta_{in} \left[\varepsilon_{n,y_i} - 1 + \frac{(\alpha'_{iy_i} - \alpha'_{in})}{\psi_{ii}^{-1}} \right] \geq \sum_{i,n} \eta_{in} \xi_i = \sum_i \xi_i \tag{15}$$

The left-hand side of Inequation (15) can be bounded by:

$$\begin{aligned} &\sum_{i,n} \eta_{in} \left[\varepsilon_{ny_i} - 1 + \frac{(\alpha'_{iy_i} - \alpha'_{in})}{\psi_{ii}^{-1}} \right] \\ &\leq \sum_i \left(\sum_n \eta_{in} \max_r \left\{ \varepsilon_{ry_i} - 1 + \frac{(\alpha'_{iy_i} - \alpha'_{ir})}{\psi_{ii}^{-1}} \right\} \right) \\ &= \sum_i \left(\sum_n \eta_{in} \bar{\xi}_i \right) = \sum_i \bar{\xi}_i \end{aligned} \tag{16}$$

□

References

[Aytar and Zisserman, 2011] Yusuf Aytar and Andrew Zisserman. Tabula rasa: Model transfer for object category detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2252–2259. IEEE, 2011.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[Cawley, 2006] Gavin C Cawley. Leave-one-out cross-validation based model selection criteria for weighted ls-svms. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pages 1661–1668. IEEE, 2006.

[Crammer and Singer, 2002] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[Gehler and Nowozin, 2009] Peter Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 221–228. IEEE, 2009.

[Griffin *et al.*, 2007] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[Jie *et al.*, 2011] Luo Jie, Tatiana Tommasi, and Barbara Caputo. Multiclass transfer learning from unconstrained priors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1863–1870. IEEE, 2011.

[Kuzborskij and Orabona, 2013] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 942–950, 2013.

[Kuzborskij *et al.*, 2013] Ilja Kuzborskij, Francesco Orabona, and Barbara Caputo. From n to n+ 1: Multiclass transfer incremental learning. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3358–3365. IEEE, 2013.

[Lampert *et al.*, 2009] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[Lu *et al.*, 2015] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14 – 23, 2015. 25th anniversary of Knowledge-Based Systems.

[Shalev-Shwartz *et al.*, 2011] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[Thrun, 1996] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? In *Advances in Neural Information Processing Systems*, pages 640–646. The MIT Press, 1996.

[Tommasi *et al.*, 2014] Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Learning categories from few examples with multi model knowledge transfer. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(5):928–941, 2014.

[Yang *et al.*, 2007] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th international conference on Multimedia*, pages 188–197. ACM, 2007.