Simple and Efficient Learning using Privileged Information

Xinxing Xu[1], **Joey Tianyi Zhou**[1], **Ivor W. Tsang**[2], **Zheng Qin**[1], **Rick Siow Mong Goh**[1] and **Yong Liu**[1]
[1] Institute of High Performance Computing, A*STAR, Singapore
[2] University of Technology Sydney
{xuxinx, zhouty, qinz, gohsm, liuyong}@ihpc.a-star.edu.sg, ivor.tsang@uts.edu.au

Abstract

The Support Vector Machine using Privileged Information (SVM+) has been proposed to train a classifier to utilize the additional privileged information that is only available in the training phase but not available in the test phase. In this work, we propose an efficient solution for SVM+ by simply utilizing the squared hinge loss instead of the hinge loss as in the existing SVM+ formulation, which interestingly leads to a dual form with less variables and in the same form with the dual of the standard SVM. The proposed algorithm is utilized to leverage the additional web knowledge that is only available during training for the image categorization tasks. The extensive experimental results on both Caltech101 and WebQueries datasets for image categorization tasks show that our proposed method can achieve a factor of up to hundred times speedup with the comparable accuracy when compared with the existing SVM+ method.

1 Introduction

In traditional machine learning paradigm, the classifiers are trained based on the features of training data, and the test is done using the same type of features. However, in real-world applications, there will be additional information that is only available during training but not available during test. For example, for the image categorization problem, we are usually classifying the images from different categories. Take the Caltech101 data set as an example, there are a total number of 101 object categories in the training set, and each of the object category has its associated descriptions, such as the textual descriptions from Wikipedia. The descriptions about each concept can be associated to each training image simply by using its label information, and hence they are only available at the training stage, but not available during the test phase. Another example is the learning from weakly labeled web images. The images in the Web are usually associated with descriptions from tags that are uploaded by the users. The tag information can be collected during training data collection. However, during test we may only have the test images that do not contain any descriptions.

The aforementioned two examples fall into the new learning paradigm of the Learning using Privileged Information (LUPI) [Vapnik and Vashist, 2009], in which the additional information is referred to as the privileged information. The Support Vector Machine using Privileged Information (SVM+) [Vapnik and Vashist, 2009] has been proposed for utilizing the privileged information. As the hinge loss is used in [Vapnik and Vashist, 2009], we refer to the formulation in their work as SVM1+ in the following. It has been proved theoretically that the incorporating of the additional privileged information can improve the convergence rate [Pechyony and Vapnik, 2010]. The SVM+ attracts much attention recently [Vapnik and Izmailov, 2015; Lopez-Paz *et al.*, 2016; Lapin *et al.*, 2014], and has been successfully applied to different applications [Niu *et al.*, 2016; Sharmanska *et al.*, 2013; Fouad *et al.*, 2012]. Following this learning scenario, some recent works proposed to utilize the privileged information for the different learning scenarios such as learning to rank [Sharmanska *et al.*, 2013], Gaussian Process classifier [Hernández-Lobato *et al.*, 2014], clustering [Feyereisl and Aickelin, 2012], distance metric learning [Xu *et al.*, 2015] and hashing [Zhou *et al.*, 2016].

However, the proposed optimization method for SVM1+ in [Vapnik and Vashist, 2009] is based on its dual form. If there are a total number of n training data, the corresponding dual problem is a Quadratic Programming (QP) problem with $2n$ variables, which is two times larger than the dual form of the standard SVM [Cortes and Vapnik, 1995] with only n variables. Besides, as the constraints in the dual form of SVM+ are different with those in the dual form of the standard SVM, the efficient implementations and the off-the-shelf solvers for SVM such as LibSVM [Chang and Lin, 2001] and SVMLight [Joachims, 1999] can not be readily utilized, and the additional efforts must be spent to design the specific solvers for optimization of SVM1+ [Pechyony *et al.*, 2010].

To overcome the aforementioned problem, in this work, we study the optimization of SVM+. Specifically, we propose a new objective function called Support Vector Machine using Privileged Information with Squared Hinge Loss (SVM2+) by simply replacing the hinge loss in SVM1+ with the squared hinge loss. The dual form of SVM2+ is a Quadratic Programming (QP) problem with only n variables, which significantly reduces the training complexity. What's more, the dual form of SVM2+ shares the same form with the dual of the standard SVM. Therefore, the existing off-the-shelf QP solvers for SVM can be readily applied to solve the proposed objective function.

We apply our proposed approach to the image categorization tasks by leveraging the additional information from web knowledge. We evaluate our algorithms on Caltech101 dataset with the additional textual descriptions for the training data obtained from Wikipedia, as well as the WebQueries dataset with the additional tag information. The extensive empirical results show that our proposed SVM2+ can achieve 110.6 (*resp.* 92.5) times speedup on Caltech101 (*resp.* WebQueries) when compared with the solution for SVM1+ as in [Vapnik and Vashist, 2009] using general QP solver yet with comparable classification accuracies. Hence, the experimental evaluation demonstrates both the efficiency and the effectiveness of our proposed SVM2+ algorithm for utilizing the additional web knowledge as privileged information for the task of image categorization.

In the following, we firstly introduce the SVM1+ in Section 2. In Section 3, we propose the objective function and the solution for SVM2+. The experimental results are shown in Section 4 and finally the conclusions are given in Section 5.

2 Support Vector Machine using Privileged Information with Hinge Loss

For the classical machine learning paradigm, we are given a set of n independent and identically distributed (IID) training pairs $\{(\mathbf{x}_1, \mathrm{y}_1), \dots, (\mathbf{x}_n, \mathrm{y}_n)\}$ with $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and $\mathrm{y}_i \in \mathcal{Y} \subset \mathbb{R}$ that are generated according to a fixed but unknown probability distribution $P(\mathbf{x}, \mathrm{y})$. For binary classification task, we have $\mathcal{Y} = \{-1, +1\}$. The target of any machine learning algorithm is therefore to learn a function $f(\mathbf{x})$ that can make the risk functional $R(f) = \frac{1}{2}\int |\mathrm{y} - f(\mathbf{x})| \mathrm{d}P(\mathbf{x}, \mathrm{y})$ minimized. Different from the classical learning paradigm, the Learning using Privileged Information (LUPI) paradigm tackles the learning senario where an additional "teacher" is only available in the training phase but not available during the test phase. Specifically, we are given a set of n independent and identically distributed training triplets $\{(\mathbf{x}_1, \mathbf{z}_1, \mathrm{y}_1), \dots, (\mathbf{x}_n, \mathbf{z}_n, \mathrm{y}_n)\}$ with $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, $\mathbf{z}_i \in \mathcal{Z} \subset \mathbb{R}^s$ and $\mathrm{y}_i \in \mathcal{Y} \subset \mathbb{R}$ that are generated according to a fixed but unknown probability distribution $P(\mathbf{x}, \mathbf{z}, \mathrm{y})$. In the test phase, the aim is the same with that of the classical learning paradigm, *i.e.*, classifying any test samples $\mathbf{x}_i$'s, where $i = n+1, \dots, n+m$. The $\mathcal{X}$ is referred to as the decision space as the test is done only in $\mathcal{X}$, while $\mathcal{Z}$ is referred to as the correcting space [Pechyony and Vapnik, 2010].

The support vector machine using privileged information (SVM+) has been proposed in [Vapnik and Vashist, 2009] for utilizing the privileged information. Specifically, $f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$ is the decision function based on the main feature $\mathbf{x}$ and $g(\mathbf{z}) = \mathbf{z}'\mathbf{v} + \rho$ is the correcting function based on the privileged information $\mathbf{z}$, and the SVM+ with hinge loss (SVM1+) is proposed to optimize the following objective function [Vapnik and Vashist, 2009]:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{v},b,\rho,\xi_i} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i + \frac{\lambda}{2}||\mathbf{v}||^2 \\ \text{s.t.,} \quad & \mathbf{z}_i'\mathbf{v} + \rho = \xi_i, \\ & \mathrm{y}_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

The objective function in (1) is solved in its dual form as in [Vapnik and Vashist, 2009; Vapnik and Izmailov, 2015]. Specifically, by introducing the non-negative multipliers $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]'$ and $\boldsymbol{\beta} = [\beta_1, \dots, \beta_n]'$, the dual form is given as in the following Quadratic Programming (QP) problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \boldsymbol{\alpha}'\mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha} \odot \mathbf{y})'\mathbf{K}(\boldsymbol{\alpha} \odot \mathbf{y}) \\ & -\frac{1}{2\lambda}(\boldsymbol{\alpha} + \boldsymbol{\beta} - C\mathbf{1})'\tilde{\mathbf{K}}(\boldsymbol{\alpha} + \boldsymbol{\beta} - C\mathbf{1}) \\ \text{s.t.,} \quad & \sum_{i=1}^{n}(\alpha_i + \beta_i - C) = 0, \\ & \sum_{i=1}^{n}\alpha_i \mathrm{y}_i = 0, \\ & \alpha_i \geq 0, \beta_i \geq 0, \end{aligned} \tag{2}$$

where $\mathbf{y} = [\mathrm{y}_1, \dots, \mathrm{y}_n]' \in \mathbb{R}^n$ is the label vector, $\mathbf{1} = [1, \dots, 1]' \in R^n$, $\odot$ is the elementwise product between two vectors/matrices, $\mathbf{K} \in \mathbb{R}^{n \times n}$ with $\mathbf{K}_{ij} = \mathbf{x}_i'\mathbf{x}_j$ is the kernel matrix obtained from the main feature $\mathbf{x}$, while $\tilde{\mathbf{K}} \in \mathbb{R}^{n \times n}$ with $\tilde{\mathbf{K}}_{ij} = \mathbf{z}_i'\mathbf{z}_j$ is the kernel matrix constructed using the privileged information $\mathbf{z}$. Although it is linear kernel here, any type of non-linear kernel [Xu *et al.*, 2013] can be readily utilized in (2).

Note that the dual problem in (2) is a standard $2n \times 2n$ QP problem, and it can be solved by the general QP solver. However, the $2n \times 2n$ problem in (2) takes more memory and requires more training time than the original SVM dual form in a $n \times n$ QP problem. Moreover, due to the introduced constraints for the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the existing efficient solvers for the SVM (*e.g.*, LibSVM and SVMLight) cannot be readily applied to solve (2). Therefore, additional efforts are required for the development of the efficient solution and the softwares such as [Pechyony *et al.*, 2010].

In the following, we propose the SVM+ with squared hinge loss, and we are interested to see that we can obtain a QP problem in size of $n \times n$ that is also in the same form with the standard dual form of SVM and can therefore be solved using any off-the-shelf efficient QP solvers developed for SVM.

3 Support Vector Machine using Privileged Information with Squared Hinge Loss

In order to solve the SVM+ more efficiently and motivated by the using of squared hinge loss for SVM [Cristianini and Shawe-Taylor, 2000], we propose the Support Vector Machine using Privileged Information with Squared Hinge Loss (SVM2+). We learn the decision function $f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$ as well as the correcting function $g(\mathbf{z}) = \mathbf{v}'\mathbf{z} + \rho$[1] by optimizing the following objective function:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{v},b,\rho,\xi_i} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 + \frac{\lambda}{2}\left(||\mathbf{v}||^2 + \rho^2\right) \\ \text{s.t.,} \quad & \mathbf{v}'\mathbf{z}_i + \rho = \xi_i, \\ & \mathrm{y}_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1 - \xi_i, \end{aligned} \tag{3}$$

where C and λ are still the two regularization parameters.

In the objective function (3), we just simply replace the original hinge loss (*i.e.*, $\sum_{i=1}^{n}\xi_i$) in (1) with the squared hinge loss (*i.e.*, $\sum_{i=1}^{n}\xi_i^2$). Besides, we also regularize the bias term ρ in the objective function as it facilitates the derivation of the dual form as shown in the proof later. The improvement looks quite straightforward, but we can observe the benefits in the dual form as shown in the following proposition 1.

Proposition 1. *The dual form of the optimization problem in (3) is given as in the following form:*

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}(\boldsymbol{\alpha}\odot\mathbf{y})'\left(\mathbf{K} + \mathbf{Q}_\lambda\odot(\mathbf{y}\mathbf{y}')\right)(\boldsymbol{\alpha}\odot\mathbf{y}) - \boldsymbol{\alpha}'\mathbf{1}, \\ s.t. \quad & \boldsymbol{\alpha}\geq 0, \boldsymbol{\alpha}'\mathbf{y} = 0, \end{aligned} \tag{4}$$

where $\mathbf{y} = [\mathrm{y}_1, \ldots, \mathrm{y}_n]' \in \mathbb{R}^n$ *is the label vector,* $\mathbf{1} = [1, \ldots, 1]' \in R^n$, $\odot$ *is the elementwise product between any two vectors/matrices,* $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]' \in \mathbb{R}^n$ *is the vector of the Lagrangian multipliers,* $\mathbf{K} \in \mathbb{R}^{n\times n}$ *with* $\mathbf{K}_{ij} = \mathbf{x}_i'\mathbf{x}_j$ *is the kernel matrix constructed from* $\mathbf{x}$*, and* $\mathbf{Q}_\lambda \in \mathbb{R}^{n\times n}$ *is a deformed kernel matrix in the form of*

$$\mathbf{Q}_\lambda = \frac{1}{\lambda}\left(\tilde{\mathbf{K}} - \tilde{\mathbf{K}}\left(\frac{\lambda}{C}\mathbf{I}_n + \tilde{\mathbf{K}}\right)^{-1}\tilde{\mathbf{K}}\right), \tag{5}$$

where $\tilde{\mathbf{K}} \in \mathbb{R}^{n\times n}$ *with* $\tilde{\mathbf{K}}_{ij} = \mathbf{z}_i'\mathbf{z}_j + 1$ *is the kernel matrix constructed from* $\mathbf{z}$*.*

Note that similarly for SVM1+, we just use the linear kernel as a demonstration, but any type of non-linear kernel can be readily utilized in (5).

Interestingly, we observe that the optimization problem in (4) is just a $n \times n$ Quadratic Programming (QP) problem. We only need to optimize with respect to the n dual variables $\boldsymbol{\alpha}$ rather than the $2n$ dual variables as in (2). The reduced dual problem can not only save memory for optimization but also reduce the computational complexity. Although there is an additional matrix inversion operation as in (5), it can be done very efficiently when compared with the optimization of the QP problem. We reduce the size of the dual form by utilizing the squared hinge loss instead of the hinge loss. By incorporating the squared hinge loss, we can easily remove the non-negative constraint for the slack variable ξ_i, and we therefore decrease the number of linear constraints in the primal form.

More importantly, the QP problem in (4) also shares the same form with the QP problem of that of the classical SVM [Cortes and Vapnik, 1995]. The difference is the changing of the kernel matrix as in the QP problem. Specifically, we can just replace the kernel matrix $\mathbf{K}$ in the original SVM to be $\mathbf{K} + \mathbf{Q}_\lambda \odot (\mathbf{y}\mathbf{y}')$. More interestingly, the matrix $\mathbf{Q}_\lambda$ is simply a transformation of the kernel matrix $\tilde{\mathbf{K}}$ constructed on the privileged information by using (5). The QP problem in (4) can be readily solved by any existing solvers (*e.g.*, LibSVM) specifically developed for the standard SVM.

We give the detailed proof of Proposition 1 as follows.

Proof. To obtain a dual form with only n variables, we must firstly get the following equivalent form for (3) by eliminating the slack variable ξ_i's:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{v},b,\rho} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_{i=1}^{n}\left(\mathbf{z}_i'\mathbf{v} + \rho\right)^2 + \frac{\lambda}{2}\left(||\mathbf{v}||^2 + \rho^2\right) \\ \text{s.t.,} \quad & \mathrm{y}_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1 - \left(\mathbf{z}_i'\mathbf{v} + \rho\right). \end{aligned} \tag{6}$$

To facilitate the derivation, let us denote $\tilde{\mathbf{v}} = [\mathbf{v}; \rho] \in \mathbb{R}^{s+1}$, and $\tilde{\mathbf{z}} = [\mathbf{z}; 1] \in \mathbb{R}^{s+1}$, and then we can optimize $\tilde{\mathbf{v}}$ for both $\mathbf{v}$ and ρ. By using these new notations, we can further obtain an equivalent optimization problem for (6) as follows:

$$\begin{aligned} \min_{\mathbf{w},\tilde{\mathbf{v}},b} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_{i=1}^{n}\left(\tilde{\mathbf{z}}_i'\tilde{\mathbf{v}}\right)^2 + \frac{\lambda}{2}||\tilde{\mathbf{v}}||^2 \\ \text{s.t.,} \quad & \mathrm{y}_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1 - \tilde{\mathbf{z}}_i'\tilde{\mathbf{v}}. \end{aligned} \tag{7}$$

Therefore, by introducing the non-negative Lagrangian multipliers α_i's only we get the Lagrangian of (6) as follows:

$$\begin{aligned} \mathcal{L} \; = \; & \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_{i=1}^{n}\left(\tilde{\mathbf{z}}_i'\tilde{\mathbf{v}}\right)^2 + \frac{\lambda}{2}||\tilde{\mathbf{v}}||^2 \\ & - \sum_{i=1}^{n}\alpha_i\left(\mathrm{y}_i\left(\mathbf{w}'\mathbf{x}_i + b\right) - 1 + \tilde{\mathbf{v}}'\tilde{\mathbf{z}}_i\right), \end{aligned} \tag{8}$$

and by setting the derivatives of $\mathcal{L}$ with respect to the primal variables $\mathbf{w}, b$ and $\tilde{\mathbf{v}}$ to be zeros, we can obtain the corresponding Karush-Kuhn-Tacker (KKT) conditions,

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n}\alpha_i\mathrm{y}_i\mathbf{x}_i = 0, \\ \frac{\partial\mathcal{L}}{\partial b} &= -\sum_{i=1}^{n}\alpha_i\mathrm{y}_i = 0, \\ \frac{\partial\mathcal{L}}{\partial\tilde{\mathbf{v}}} &= \lambda\tilde{\mathbf{v}} - \sum_{i=1}^{n}\alpha_i\tilde{\mathbf{z}}_i + C\sum_{i=1}^{n}\left(\tilde{\mathbf{v}}'\tilde{\mathbf{z}}_i\right)\tilde{\mathbf{z}}_i = 0, \end{aligned}$$

which further lead to the following:

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i\mathrm{y}_i\mathbf{x}_i, \tag{9}$$

$$\sum_{i=1}^{n}\alpha_i\mathrm{y}_i = 0, \tag{10}$$

$$\tilde{\mathbf{v}} = \left(\lambda\mathbf{I}_{(s+1)} + C\sum_{i=1}^{n}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i'\right)^{-1}\sum_{i=1}^{n}\alpha_i\tilde{\mathbf{z}}_i. \tag{11}$$

[1] Recently, the correcting function with label information has been utilized in [Vapnik and Izmailov, 2015] (*i.e.*, $g(\mathbf{z}) = \mathrm{y}(\mathbf{v}'\mathbf{z} + \rho)$), which can also be easily incorporated in our formulation.

Algorithm 1 The Fast Solution for SVM2+

Calculate the kernel matrix $\mathbf{K}$ and $\tilde{\mathbf{K}}$ using $\mathbf{x}$ and $\mathbf{z}$.
Obtain the deformed kernel matrix $\mathbf{Q}_\lambda$ using (5).
Obtain the dual variables $\boldsymbol{\alpha}$ using SVM solver (*e.g.*, LibSVM).
return $f(\mathbf{x})$.

To obtain a dual form with kernelization [Boser *et al.*, 1992], let us denote $\sum_{i=1}^{n} \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i' = \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}'$ with $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_n] \in \mathbb{R}^{(s+1)\times n}$, and according to Woodbury matrix identity[2], we have:

$$\left(\lambda \mathbf{I}_{(s+1)} + C \sum_{i=1}^{n} \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i'\right)^{-1} = \frac{1}{\lambda}\left(\mathbf{I}_{(s+1)} - \tilde{\mathbf{Z}}\left(\frac{\lambda}{C}\mathbf{I}_n + \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\right)^{-1}\tilde{\mathbf{Z}}'\right). \quad (12)$$

By substituting the (9), (10) and (11) as well as (12) back into the Lagrangian (8), we can finally obtain

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \mathrm{y}_i \alpha_j \mathrm{y}_j \mathbf{x}_i' \mathbf{x}_j - \frac{1}{2\lambda}\boldsymbol{\alpha}'\tilde{\mathbf{Z}}'\left(\mathbf{I}_{(s+1)} - \tilde{\mathbf{Z}}\left(\frac{\lambda}{C}\mathbf{I}_n + \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\right)^{-1}\tilde{\mathbf{Z}}'\right)\tilde{\mathbf{Z}}\boldsymbol{\alpha}, \quad (13)$$

and by denoting $\mathbf{K} = \mathbf{X}'\mathbf{X}$, $\tilde{\mathbf{K}} = \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}$ and $\mathbf{Q}_\lambda = \frac{1}{\lambda}\left(\tilde{\mathbf{K}} - \tilde{\mathbf{K}}\left(\frac{\lambda}{C}\mathbf{I}_n + \tilde{\mathbf{K}}\right)^{-1}\tilde{\mathbf{K}}\right)$ as a deformed kernel matrix, we get the dual form as in (4). Thus we finish the proof. □

After solving the dual form as in (4), we obtain the final decision function as

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \mathrm{y}_i \mathbf{x}_i'\mathbf{x} + b = \sum_{i=1}^{n} \alpha_i \mathrm{y}_i \mathbf{K}(\mathbf{x}, \mathbf{x}_i) + b. \quad (14)$$

Therefore, our proposed algorithm for solving SVM2+ is summarized as in Algorithm 1.

4 Experiments

In this section, we show the experimental results of our proposed algorithm and the baseline algorithms for image categorization tasks on the Caltech101 [Li *et al.*, 2007] and the WebQueries [Krapac *et al.*, 2010] datasets.

We compare our proposed new algorithm SVM2+ with the baseline algorithms, *i.e.*, SVM, SVM-2K [Farquhar *et al.*, 2005] and SVM1+. The SVM is trained based only on the visual feature extracted from images, and the SVM2K is a two-view learning algorithm that trains two classifiers on the two views simultaneously, and we only use the view from visual feature for prediction. Therefore, the SVM2K, SVM1+ and SVM2+ all utilized the additional privileged information during training. However, only the images are used for the test for all the algorithms as the privileged information is not available in our learning setting.

[2] $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$

4.1 Dataset Description and Feature Extraction

Table 1 summarizes the details of the used two datasets for our experiments, which are described in the following.

Caltech101

The Caltech101 dataset[3] contains images from 101 object categories (*e.g.*, “helicopter”, “elephant” and “chair” etc.) and a background category that contains the images not from the 101 object categories. For each object category, there are about 40 to 800 images, while most classes have about 50 images. The resolution of the image is roughly about 300×200 pixels. Following the popular settings [Li *et al.*, 2007], we utilize 10 images per class for training and up to 50 images per category for the test. Finally, we get 1020 images in the training set, and 2995 images in the test set.

WebQueries

The WebQueries dataset[4] is composed of 71,478 images obtained by retrieving a total number of 353 textual web queries (*e.g.*,“eiffel tower”, “violin” and “France flag” etc.). For each image in the WebQueries dataset, there are corresponding textual descriptions either in English or French. Besides, the relevant labels have been annotated by human manually. In our experiments, the images with English queries and the textual queries with more than 100 images according to the ground truth labels are used as the evaluation queries. In this way, we obtain a total number of 76 queries for the final classification tasks. For each of the 76 queries, we used 10 (*resp.*, 50) images for constructing the training set (*resp.*, test set). Therefore, we have 760 images for training, while we have 3800 images for test.

Image Feature Extraction

For the image representation, the deep learning features are extracted from each of the images due to its excellent performance for computer vision tasks [Donahue *et al.*, 2014]. The MatConvNet [Vedaldi and Lenc, 2015] is used to extract the deep learning features. The vgg-s model [Chatfield *et al.*, 2014] is used in our work. It is pre-trained on the 1.2 million ImageNet dataset [Deng *et al.*, 2009], and the 4096-dimensional output of the fc6 in the deep Convolutional Neural Network (CNN) model from each image is employed as the visual feature representation. The same type of visual features are extracted for both the Caltech101 and the WebQueries datasets.

Web Knowledge as Privileged Information

For the Caltech101, it is difficult to obtain descriptions for each of the image. We therefore discover the web knowledge by searching the category name of each category from Wikipedia. Then we collect the text descriptions of each concept from the webpage of Wikipedia. Obtaining the textual descriptions for all the categories, we further use the term frequency-inverse document frequency (TF-IDF) to convert each textual description into the bag of word frequency feature vector. We finally get a 29,535 dimensional feature vector for each of the object category. During training, each

[3] http://www.vision.caltech.edu/Image_Datasets/Caltech101

[4] https://lear.inrialpes.fr/~krapac/webqueries/webqueries.html

Table 1: A summarization of the datasets used in our experiments. The n and m are the number of training images and test images, and #c is the number of total classes, and d and s are the feature dimensions for the main feature and privileged feature, respectively.

	n	m	#c	d	s
Caltech101	1020	2995	102	4096	29535
WebQueries	760	3800	76	4096	2000

Table 2: The classification accuracies (%) of all methods on the Caltech101 and the WebQueries datasets.

	Caltech101	WebQueries
SVM	85.41	53.74
SVM2K	84.82	53.21
SVM1+	85.79	54.30
SVM2+	86.20	54.43

training image is associated with one vector that represents its object category as its privileged information. As we do not have the ground-truth labels for test images, the privileged information is not available during the test phase.

For the WebQueries dataset, each image is associated with a tag that contains a short description of the image in the Website. We also collect the textual descriptions from each of the training images, and then we remove the stop-words and calculate the term-frequency (TF) for each of the textual descriptions. The top-2000 words from the whole corpora are used as the vocabulary, and finally each textual description is represented as a 2000-dimensional feature vector.

4.2 Experimental Results and Discussion

We use the linear kernel for both the visual feature and the privileged textual feature to train one-vs-others classifiers for each object category or query, and then we assign the labels of the test image to be the one with the highest output from classifiers on all categories or queries. The classification accuracy is used as the performance measurement.

The experimental results with the classification accuracies for the different algorithms are shown as in Table 2. We can observe that the additional privileged information does help the classification tasks for both the datasets by using the LUPI framework. Therefore, it is beneficial to utilize the web knowledge as privileged information for the task of image categorization. Besides, the SVM2K is worse than SVM and the other algorithms, which shows that the training of the two-view classifiers is not effective for the LUPI learning paradigm. We observe that the results from SVM2+ on the two datasets are slightly better than the results from SVM1+, which demonstrates that it may be more suitable to utilize the squared hinge loss for utilizing the privileged information on these two applications. In general, using the hinge loss or squared hinge loss will lead to comparable performances for different tasks as demonstrated in [Lee and Lin, 2013]. However, we are using the squared hinge loss to speed up the training phase of the SVM+ as shown later.

We compare the training CPU time of the different al-

Table 3: The training CPU time (Seconds) of all methods on the Caltech101 and WebQueries datasets.

	Caltech101	WebQueries
SVM	17.86	1.997
SVM2K	732.8	80.79
SVM1+	867.6	189.0
SVM2+	7.847	2.044

Table 4: The relative speedup times of our proposed SVM2+ with respect to SVM1+ on the Caltech101 and WebQueries datasets.

	Caltech101	WebQueries
Speedup times	110.6	92.5

gorithms with a Lenovo desktop (3.20GHz CPU with 8GB RAM) and Matlab implementation. The parameters with the best classification accuracy for each algorithm are used to report the training time. As the SVM1+ cannot be fit into the standard QP solver for SVM, we utilize the commercial software Mosek[5] with academic license to solve the Quadratic Programming problem. For SVM and SVM2+, we all utilize the LibSVM[6] solver. The results in Table 3 show the CPU time of each algorithm on the whole Caltech101 and WebQueries datasets. Table 4 shows the speedup times of our proposed SVM2+ when compared with the SVM1+. We can observe from the table that our proposed SVM2+ achieves 110.6 and 92.5 times speedup respectively on Caltech101 and WebQueries datasets when compared with SVM1+.

5 Conclusions

In this work, we propose a simple but effective SVM2+ objective function by utilizing the squared hinge loss instead of the hinge loss as in the traditional SVM1+ method, which leads to up to hundred times speedup for training the SVM+ classifiers on the image categorization datasets. In the future, we would like to study the convergence rate and the theoretical bound for our proposed SVM2+.

References

[Boser *et al.*, 1992] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992.*, pages 144–152, 1992.

[Chang and Lin, 2001] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM:a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[Chatfield *et al.*, 2014] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details:

[5] https://www.mosek.com/

[6] https://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[Cristianini and Shawe-Taylor, 2000] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Donahue *et al.*, 2014] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[Farquhar *et al.*, 2005] Jason D. R. Farquhar, David R. Hardoon, Hongying Meng, John Shawe-Taylor, and Sándor Szedmák. Two view learning: Svm-2k, theory and practice. In *NIPS*, pages 355–362, 2005.

[Feyereisl and Aickelin, 2012] Jan Feyereisl and Uwe Aickelin. Privileged information for data clustering. *Inf. Sci.*, 194:4–23, 2012.

[Fouad *et al.*, 2012] Shereen Fouad, Peter Tiño, Somak Raychaudhury, and Petra Schneider. Learning using privileged information in prototype based models. In *ICANN*, pages 322–329, 2012.

[Hernández-Lobato *et al.*, 2014] Daniel Hernández-Lobato, Viktoriia Sharmanska, Kristian Kersting, Christoph H. Lampert, and Novi Quadrianto. Mind the nuisance: Gaussian process classification using privileged noise. In *NIPS*, pages 837–845, 2014.

[Joachims, 1999] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[Krapac *et al.*, 2010] Josip Krapac, Moray Allan, Jakob J. Verbeek, and Frédéric Jurie. Improving web image search results using query-relative classifiers. In *CVPR*, pages 1094–1101, 2010.

[Lapin *et al.*, 2014] Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: SVM+ and weighted SVM. *Neural Networks*, 53:95–108, 2014.

[Lee and Lin, 2013] Ching-Pei Lee and Chih-Jen Lin. A study on l2-loss (squared hinge-loss) multiclass SVM. *Neural Computation*, 25(5):1302–1323, 2013.

[Li *et al.*, 2007] Fei-Fei Li, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[Lopez-Paz *et al.*, 2016] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. ICLR, 2016.

[Niu *et al.*, 2016] Li Niu, Xinxing Xu, Lin Chen, Lixin Duan, and Dong Xu. Action and event recognition in videos by learning from heterogeneous web sources. *IEEE Trans. Neural Netw. Learning Syst.*, March 2016.

[Pechyony and Vapnik, 2010] Dmitry Pechyony and Vladimir Vapnik. On the theory of learnining with privileged information. In *NIPS*, pages 1894–1902, 2010.

[Pechyony *et al.*, 2010] Dmitry Pechyony, Rauf Izmailov, Akshay Vashist, and Vladimir Vapnik. Smo-style algorithms for learning using privileged information. In *DMIN*, pages 235–241, 2010.

[Sharmanska *et al.*, 2013] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H. Lampert. Learning to rank using privileged information. In *Proceedings of the 14th International Conference on Computer Vision*, pages 825–832, Sydney, Australia, Dec. 2013.

[Vapnik and Izmailov, 2015] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.

[Vapnik and Vashist, 2009] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

[Vedaldi and Lenc, 2015] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for MATLAB. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 689–692, 2015.

[Xu *et al.*, 2013] Xinxing Xu, Ivor W. Tsang, and Dong Xu. Soft margin multiple kernel learning. *IEEE Trans. Neural Netw. Learning Syst.*, 24(5):749–761, 2013.

[Xu *et al.*, 2015] Xinxing Xu, Wen Li, and Dong Xu. Distance metric learning using privileged information for face verification and person re-identification. *IEEE Trans. Neural Netw. Learning Syst.*, 26(12):3150–3162, 2015.

[Zhou *et al.*, 2016] Joey Tianyi Zhou, Xinxing Xu, Sinno Jialin Pan, Ivor W. Tsang, Zheng Qin, and Rick Siow Mong Goh. Transfer hashing with privileged information. In *IJCAI*, 2016.