Learning from the Mistakes of Others: Matching Errors in Cross-Dataset Learning

Viktoriia Sharmanska and Novi Quadrianto
SMiLe CLiNiC, University of Sussex, Brighton, UK
sharmanska.v@gmail.com; n.quadrianto@sussex.ac.uk

Abstract

Can we learn about object classes in images by looking at a collection of relevant 3D models? Or if we want to learn about human (inter-)actions in images, can we benefit from videos or abstract illustrations that show these actions? A common aspect of these settings is the availability of additional or privileged data that can be exploited at training time and that will not be available and not of interest at test time. We seek to generalize the learning with privileged information (LUPI) framework, which requires additional information to be defined per image, to the setting where additional information is a data collection about the task of interest. Our framework minimizes the distribution mismatch between errors made in images and in privileged data. The proposed method is tested on four publicly available datasets: Image+ClipArt, Image+3Dobject, and Image+Video. Experimental results reveal that our new LUPI paradigm naturally addresses the cross-dataset learning.

1. Introduction

Vapnik et al. [35, 24, 34] introduced learning with privileged information (LUPI) as a learning with teacher paradigm, where at the training stage, a teacher gives some additional *explanation* $\mathbf{x}_i^\star$ about an example $\mathbf{x}_i$. LUPI has been shown useful in a variety of learning scenarios such as ranking [28], categorization [37], structured prediction [9], data clustering [8], metric learning [10], face/gesture recognition [38], glaucoma detection [7], and recently learning with annotation disagreements [27]. Most LUPI methods (e.g. [35, 28, 20, 14]) follow the assumption that the extra information is useful to discriminate between easy and difficult examples. This knowledge is then used to determine the influence of each instance in the training process. Specifically, one puts less emphasis or even ignores difficult instances during training in hope that this will improve performance. Reflecting on how per-instance privileged information can be used to identify whether this instance is an easy or a difficult one, we ask the following question: is it possible to transfer easiness and hardness in *a class sense* instead of *per-instance*?

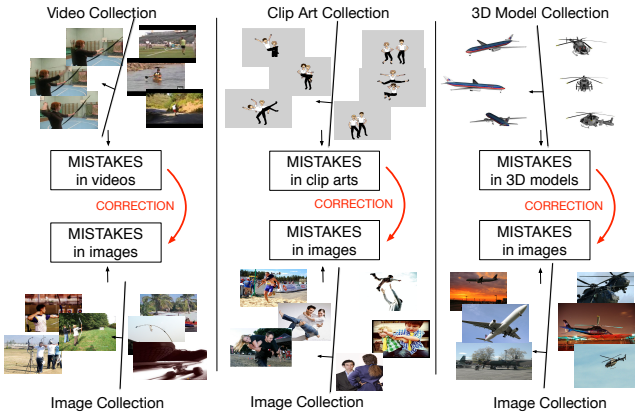


Figure 1: In this work, we propose a framework to solve vision tasks *in images* by acquiring knowledge from the mistakes committed by other data collections (videos, clip arts, and 3D models) when learning the same concepts.

This paper seeks to advice the learner to acquire knowledge from the mistakes committed by others when learning the same concept. One will learn that making errors on an example-by-example basis is unavoidable, but one will make the right decisions at a larger scale by minimizing divergence between own mistakes and others'. We will explore a distribution matching over the mistakes in original and privileged representations as a principled approach to achieve the class-level information transfer. We will use the regularized risk functional framework and replace the empirical risk with an (empirical) divergence term characterizing the mismatch between error distributions on the privileged and original spaces. This approach is innovative in two senses: **(1)** Prior knowledge is normally encoded in the regularization term, instead we introduce bias into the risk (loss) term. **(2)** Almost all distribution matching methods match input features and/or function outputs, instead we match error distributions.

2. Related work

In the literature, distribution matching has been proposed for, among others, transduction learning (e.g. [25]) and domain adaptation (e.g. [21, 40]). Quadrianto et al. [25] matched the distribution between function outputs on the training data $f(X) := \{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)\}$ and outputs on the test data $f(X') := \{f(\mathbf{x}'_1), \ldots, f(\mathbf{x}'_{N'})\}$ to devise a general transduction algorithm for classification, regression, and structured prediction settings, whereas we propose to match *error* distributions on privileged and original domains. The empirical Maximum Mean Discrepancy (MMD) [12] is employed as the nonparametric metric of difference between two distributions. In the domain adaptation setting, Pan et al. [23] used the MMD metric to project data from a target domain $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and a related source domain $X' := \{\mathbf{x}'_1, \ldots, \mathbf{x}'_{N'}\}$ into a common subspace such that the difference between the distributions of source and target domains is reduced. Recently, Zhang et al. [40] proposed to also use the MMD metric to project data X and X' as well as function outputs $f(X)$ and $f(X')$ in the framework of deep neural networks.

In general, finding projection matrices involves either transformation of the data from source and target into a common subspace (two projection matrices) or transformation of the data from source to target (one projection matrix). The projection methods can be expensive in both computational complexity and memory requirement (if the data dimensionality is high). Our method offers a refreshing look on domain adaptation problems that sidestep the process of finding projection matrices. The cross-dataset scenario in this paper overlaps with the work of, for example, [32], which aims to overcome dataset bias across multiple image datasets in the domain adaptation scenario. In contrast, we explore cross-modal transfer in the cross-dataset learning. Complementary to us, [13] recently use a distillation framework for cross-modal representation learning.

Model *distillation* or compression [15, 2] has attracted attention in the domain adaptation setting with deep architectures (e.g. [33, 13]). The aim is to learn representations for an unlabeled or sparsely labeled target domain by using a large labeled source domain as a supervisory signal. The framework uses output probability predictions on source domain as training labels for target domain and shows that this training is more accurate than using the original target labels. Instead we propose to match error distributions across domains as knowledge distillation in a LUPI framework. This is also supported by a work of [22] that connects distillation and LUPI with one-to-one correspondences. In Section 3, we will describe related work on LUPI, followed by our proposed generalization of the LUPI paradigm in Section 4. We choose to motivate our work in terms of LUPI as it offers a unified framework for learning with additional information that is only available at training time.

3. LUPI with one-to-one correspondence

We formalize the LUPI setup for the task of supervised binary classification with a single source of privileged information. Assume that we are given a set of N training examples, represented by feature vectors $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{X} = \mathbb{R}^d$, their label annotations, $Y = \{y_1, \ldots, y_N\} \in \mathcal{Y} = \{+1, -1\}$, and additional information, also in the form of feature vectors, $X^\star = \{\mathbf{x}^\star_1, \ldots, \mathbf{x}^\star_N\} \subset \mathcal{X}^\star = \mathbb{R}^{d^\star}$, where $\mathbf{x}^\star_i$ encodes the additional information we have about sample $\mathbf{x}_i$. This additional information is only available at training time, thus is referred as the privileged information. We now have $\mathcal{X}$ as the original data space and $\mathcal{X}^\star$ as the privileged data space. What we want is to learn a binary classification function $f : \mathcal{X} \to \mathcal{Y}$ from a large space of possible functions $\mathcal{F}$ that can then be used to infer the label y_{new} for a new input instance $\mathbf{x}_{\text{new}}$. The goal of LUPI is to exploit the privileged information in the learning process of the latent function f. This exploitation, however, should not involve the usage of $X^\star$ information as a direct input to the function f, because $X^\star$ is not available for yet to be seen instances.

For this, a common approach found in the literature is to consider that the privileged information can be used to distinguish between *easy* and *difficult* instances [35, 28, 14]. This extra knowledge can be used to bias the learning process towards finding a latent function f with better generalization properties.

Slack based methods. Vapnik and Vashist [35] introduced an SVM+ method as a generalization of the SVM-based framework to solve LUPI. SVM+ tries to upper bound the mistake at i-th data point in the original space $\mathcal{X}$ based on the privileged data $\mathcal{X}^\star$. Intuitively, we try to predict the difficulty of each data point based on the additional privileged data for that particular instance, thereby creating a *data dependent upper bound* ξ_i on the hinge loss. In the context of binary classification with a *linear classifier*, $f(\mathbf{x}; \mathbf{w}) := \langle \mathbf{w}, \mathbf{x} \rangle + b$, the SVM+ optimization admits the following form:

$$\underset{\mathbf{w},\mathbf{w}^\star,b,b^\star}{\text{minimize}} \quad \underbrace{\|\mathbf{w}\|^2}_{\text{regularization}} + C^\star \underbrace{\|\mathbf{w}^\star\|^2}_{\text{regularization}} + C \underbrace{\sum_{i=1}^{N} [\langle \mathbf{w}^\star, \mathbf{x}^\star_i \rangle + b^\star]}_{\text{loss:=upper bound of own mistakes}} \tag{1a}$$

$$\text{subject to, for all } i = 1, \ldots, N; \ \langle \mathbf{w}^\star, \mathbf{x}^\star_i \rangle + b^\star \geq 0,$$
$$\underbrace{1 - y_i[\langle \mathbf{w}, \mathbf{x}_i \rangle + b]}_{\text{own mistake}} \leq \underbrace{\langle \mathbf{w}^\star, \mathbf{x}^\star_i \rangle + b^\star}_{\text{data dependent upper bound}}. \tag{1b}$$

SVM+ parameterizes the slack value for each sample $\xi_i = \langle \mathbf{w}^\star, \mathbf{x}^\star_i \rangle + b^\star$ with unknown $\mathbf{w}^\star$ and $b^\star$ parameters. These slack variables indicate which instances are *easy* and which are *difficult* to classify based on privileged information. Specifically, a difficult instance $\mathbf{x}^\star_i$ has a large slack

variable ξ_i, which makes the corresponding constraint in (1b) have little impact or none at all in the estimation of $\mathbf{w}$. If an instance is easy, its slack variable is close to zero, and the optimization task would concentrate on satisfying the corresponding constraint in (1b).

Remark In a variant of SVM+, called dSVM+ [35], Vapnik and Vashist first train a standard SVM parameterized by $\hat{\mathbf{w}}$ and $\hat{b}$ on $X^\star$, Y and then compute *deviation* values $d_i^\star$ of each training instance. These are defined as $d_i^\star = 1 - y_i[\langle \hat{\mathbf{w}}, \mathbf{x}_i^\star \rangle + \hat{b}]$. Finally, an SVM+ is trained using X, $X^\star$ and Y, where $X^\star$ is a column vector with i-th element that corresponds to the deviation value of i-th training instance, $d_i^\star$. This means the constraint in Eq. (1b) for a data point $\mathbf{x}_i$ is now upper bounded by a scaled and shifted version of other's mistake:

$$\underbrace{1 - y_i[\langle \mathbf{w}, \mathbf{x}_i \rangle + b]}_{\text{own mistake}} \leq \langle \mathbf{w}^\star, \underbrace{(1 - y_i[\langle \hat{\mathbf{w}}, \mathbf{x}_i^\star \rangle + \hat{b}])}_{\text{other's mistake}} \rangle + b^\star.$$

The idea is that if it is difficult to classify the instance in the privileged domain $X^\star$, which is often assumed to have better quality instances [35, 28], then it is going to be even more difficult in the original X domain.

Non-slack based methods. In an ensemble approach, Chen et al. [4] described an Adaboost algorithm that uses privileged information. The method proposed considers decision stumps as weak classifiers, which are trained on the union of X and $X^\star$ at each step. In the context of Gaussian process classification, Hernández-Lobato et al. [14] proposed a heteroscedastic Gaussian process classification model to address classification tasks with privileged information. Wang et al. [36] proposed to solve a joint regularized risk functional over f and $f^\star$ with an extra regularization term that couples the optimization problem in the form of $\sum_{i=1}^{N}(f(\mathbf{x}_i) - f^\star(\mathbf{x}_i^\star))^2$. This extra term is similar to the squared difference term in the co-regularization based multi view semi-supervised learning approaches (e.g. [3, 29]). The crucial difference is while co-regularization methods aim to improve the average performance of all single view classifiers, LUPI method is only interested in improving the performance of the classifier in the original space.

4. Correspondence-free LUPI

We seek to relax the one-to-one correspondence assumption in the standard LUPI formulation. This is natural in the setting where we learn to recognize, for example, activities in images while the same activities are also available in *videos* as privileged information. We will of course not expect that there will be a one-to-one correspondence between images and videos. Nevertheless, learning about activities from videos could be informative about the action class and applicable to the same task with images. Another example is learning about interactions among people like *dancing* from real images and from abstract illustrations. Albeit the action does not appear the same way in abstract and real images, both representations are informative about the action and can learn from one another [1].

We argue that the LUPI framework can be generalized to such scenario by transferring the general class characteristic from the privileged to the original data via the distribution of the easy and hard samples. In the following we explain the idea of matching the distribution of slack variables as a model for the class-level information transfer.

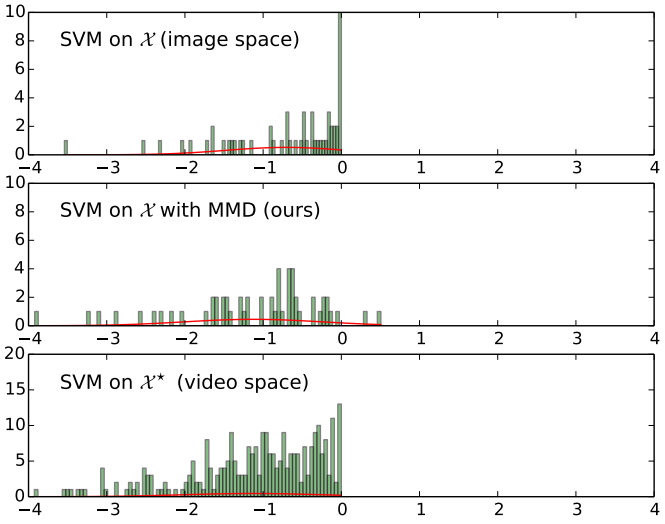
4.1. Distribution matching

We assume a linear form of classifier in the privileged and original spaces. Let $p_{d^\star}$ denote a distribution over deviation values $d^\star$ (i.e. unthresholded slack variables but we might refer them simply as slack variables if the context is clear). The deviation values $d^\star$, as in the dSVM+ formulation, are obtained by first training a linear SVM on $X^\star$, $Y^\star$ and then evaluating the slack variables using training data. The $N^\star$ samples from this distribution are denoted as $D^\star = \{d_i^\star \mid d_i^\star = 1 - y_i^\star \langle \mathbf{w}^\star, \mathbf{x}_i^\star \rangle, i = 1, \ldots, N^\star\}$. The shape of $p_{d^\star}$ reflects the distribution over easy and hard instances of the class in the privileged space. This distribution is *class specific* and can be seen as error distribution of the classifier in the privileged space. We assume that for a particular classification task like differentiating human actions, this error distribution stays similar across modalities. So that a distribution p_d over deviation values in the original space X with N samples denoted as $D = \{d_i \mid d_i = 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle,\ i = 1, \ldots, N\}$) and $p_{d^\star}$ coincide. Therefore, our main assumption is that the error distributions, $p_{d^\star}$ and p_d, could be matched even if f and $f^\star$ are learned from different modalities, images and videos, respectively.

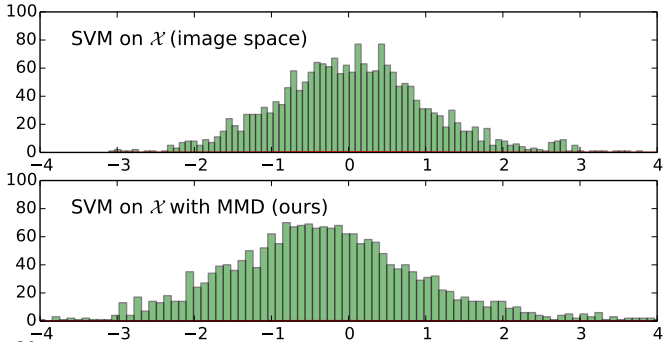
Our main objective for the *class-level* information transfer is based on the regularized risk minimization framework with a divergence term characterizing the mismatch between the class error distributions in the privileged and original spaces acting as a *loss* term:

$$\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \quad \underbrace{\|\mathbf{w}\|^2}_{\text{regularization}} + C \quad \underbrace{\text{KL}(p_{d^\star} || p_d)}_{\substack{\text{loss := divergence between} \\ \text{own mistakes and others' mistakes}}} \tag{2}$$

where $\text{KL}(p_{d^\star} || p_d)$ is the Kullback-Leibler divergence between distributions $p_{d^\star}$ and p_d and C is the trade-off hyper-parameter that controls the relative influence of the divergence (loss) component and the regularization. Note that the KL divergence is asymmetric, the choice in expressing the distribution distance measure as $\text{KL}(p_{d^\star} || p_d)$ instead of $\text{KL}(p_d || p_{d^\star})$ is deliberate. This will simplify our learning algorithm as it will become clear below. Contrasting our proposed method with SVM+, we notice that in SVM+, Eq. (1), the training error is upper bounded per each in-



(a) Training error distributions



(b) Test error distributions

Figure 2: Visualization of the error distributions of three classifiers in the experiment with Image+Video dataset. For a binary problem of differentiating a *kayaking* action from others, we compare p_d when training SVM on original data space $\mathcal{X}$, $p_{d \text{ with KL}(p_{d^\star}||p_d)}$ when training our proposed SVM MMD on data from $\mathcal{X}$ and $\mathcal{X}^\star$, and $p_{d^\star}$ when training SVM on privileged data space $\mathcal{X}^\star$. In this case, our proposed method successfully utilizes privileged information: the peak of train (2a: middle) and test (2b: middle) distributions $p_{d \text{ with KL}(p_{d^\star}||p_d)}$ are shifted to the *left* comparing to p_d. Train and test cases of $p_{d^\star}$ are the same as we use all available data *in the other dataset* as privileged information.

stance based on its privileged data (requires one-to-one correspondence). Instead in Eq. (2), we match the distribution of errors in privileged and original spaces (correspondence-free setting). Our intuition is that making errors on an instance basis is unavoidable, but we will make better decisions at a large scale by comparing error distributions.

To compute the KL divergence, we require a parametric assumption on the distribution $p_{d^\star}$ as well as p_d. If we assume that $X^\star$ is of much better quality than X as in for example [35, 28], the distribution of the slack variables on privileged space $d^\star$ will have a mean value in the negative region (for a correct prediction with a high confidence, the functional margin $y_i^\star \langle \mathbf{w}^\star, \mathbf{x}_i^\star \rangle$ will be large and therefore the slack variable will be negative). Whereas, the distribution of the slack variables in the original space, p_d, will have a mean value around zero and tails that accounts for correctly (left tail) classified samples at negative region and incorrectly (right tail) classified samples at positive region. Please, refer to our visualization of the distribution over the deviation values in Figure 2.

In the simplest case, we could model the p_d distribution with the Gaussian exponential family, $p_d = \mathcal{N}(d|\mu_d, \sigma_d^2)$. With this assumption, minimizing $\text{KL}(p_{d^\star}||p_d)$ reduces to *matching* the first and second *moments* of the two distributions, which are the mean and the variance.

4.2. Maximum Mean Discrepancy

We can go beyond the Gaussian assumption and match skewness, kurtosis (third and fourth moments) or even higher order moments. In a more general case, to avoid a parametric assumption on the distance estimate between distributions, we propose to use the Maximum Mean Discrepancy (MMD) criterion [12], a non-parametric distance estimate. Denote by $\mathcal{H}$ a Reproducing Kernel Hilbert Space with kernel k defined on $\mathcal{X}$. In this case one can show [12] that whenever k is characteristic (or universal), the map

$$\mu : p \to \mu[p] := \mathbb{E}_{d \sim p_d}[k(d, \cdot)]$$
$$\text{with associated distance}$$
$$\text{MMD}(p_{d^\star}, p_d) := \|\mu[p_{d^\star}] - \mu[p_d]\|^2 \qquad (3)$$

characterizes a distribution uniquely. Examples of characteristic kernels [31] are Gaussian RBF, Laplacian and B_{2n+1}-splines. With a this choice of kernel functions, the MMD criterion matches infinitely many moments in the Reproducing Kernel Hilbert Space (RKHS). The mean and variance matching described in the previous section is a special case when we use a polynomial kernel with degree 2. We use a biased estimate of MMD as follows:

$$\widehat{\text{MMD}} = \frac{1}{N^2} \sum_{i}^{N} \sum_{i'}^{N} k(d_i, d_{i'}) - \frac{2}{NN^\star} \sum_{i}^{N} \sum_{j}^{N^\star} k(d_i, d_j^\star) +$$
$$+ \frac{1}{N^{\star,2}} \sum_{j}^{N^\star} \sum_{j'}^{N^\star} k(d_j^\star, d_{j'}^\star). \qquad (4)$$

The above quantity is then used as a plug-in estimator for non-parametric $\text{KL}(p_{d^\star}||p_d)$ in (2). Please refer to Alg. 1 for the summary of our proposed method.

Remark Using non universal kernels such as a polynomial kernel will only give necessary but not sufficient conditions for distribution matching. Hence we use RBF kernel

Algorithm 1 Matching Error distributions on $\mathcal{X}^\star$ and $\mathcal{X}$

Input original data (X, Y), privileged data $(X^\star, Y^\star)$,
assume $f(\mathbf{x}) := \langle \mathbf{w}, \mathbf{x} \rangle$ and $f^\star(\mathbf{x}^\star) := \langle \mathbf{w}^\star, \mathbf{x}^\star \rangle$
$\mathbf{w}^\star \leftarrow$ solve $||\mathbf{w}^\star||^2$ + hingeloss$(\mathbf{w}^\star | X^\star, Y^\star)$
$D^* = \{1 - y_i^\star \langle \mathbf{w}^\star, \mathbf{x}_i^\star \rangle\}_{i=1}^{N^\star}$ (errors on $X^\star$)
$\mathbf{w} \leftarrow$ solve $||\mathbf{w}||^2$ + $\widehat{\text{MMD}}$loss $(D^\star, D(\mathbf{w}) | X, Y)$
s.t. $D(\mathbf{w}) = \{1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\}_{i=1}^{N}$ (errors on X)
Return $\mathbf{w}$

in the MMD criterion, while maintaining a linear classifier form in the proposed method. Computing MMD criterion in Eq. (4) costs $O((N+N^\star)^2)$ time [12], this is true for any kernel. We plan to explore advancements in fast two-sample test with cost that is linear in number of samples (e.g. [5]).

5. Experiments

We study the task of object as well as action recognition in images with three possible types of privileged information available at training time: *clip art illustrations*, *videos*, and *3D models*. The classification task is the same in both modalities, so that the privileged data is informative about the objects/actions that we are primarily interested to recognize using the image modality.

Datasets. We use four publicly available datasets to test the performance of our cross-modal/dataset scenario: the INTERACT[1] dataset [1] with clip art illustrations collected in addition to images that capture the interaction between people, the UCF101[2] action recognition dataset of videos [30], and the CrossLink[3] dataset [16] of 3DWarehouse[4] models accompanied by the action and object images from the ImageNet dataset[5] [26].

Methods. We compare our SVM MMD model (`SVM MMD`) with a standard object classification baseline such as SVM classifier trained on the image space $\mathcal{X}$ (`SVM Images`). To put our method into perspective of domain adaptation and provided that the feature dimensionality is the same across modalities $\mathcal{X}$ and $\mathcal{X}^\star$, we compare the proposed `SVM MMD` with the *instance-transfer* approach that shares the data samples between the two modalities, i.e. SVM trained on union of image and privileged data (`SVM Combined`); and the *model-transfer* method that relies on parameter transfer from privileged (source) to image (target) space, such as adaptive SVM [19, 39] (`SVM Adaptive`). For a given solution of the source task, $\mathbf{w}^{\text{source}}$, and training data of the target task, `SVM Adaptive` solves the following optimization problem:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \|\mathbf{w} - \mathbf{w}^{\text{source}}\|^2 + \frac{C}{N} \sum_{j=1}^{N} \xi_j \qquad (5)$$
$$\text{s.t.} \;\; 1 - y_j \langle \mathbf{w}, \mathbf{x}_j \rangle \leq \xi_j, \;\; \xi_j \geq 0 \;\; \text{for all} \;\; 1 \leq j \leq N.$$

To train a classifier on image data, we solve (5) using as $\mathbf{w}^{\text{source}}$ the weight vector obtained from training using the privileged data. From the perspective of domain adaptation, `SVM Adaptive` transfers the information by introducing the *bias into the regularization* term of SVM, whereas the proposed MMD model introduces the *bias into the loss* term of the SVM. We also provide a reference comparison with the SVM+ baseline (`SVM+`) [35] that relies on the one-to-one correspondence between samples in the original and privileged spaces if applicable (Section 5.1).

Model selection. We perform a cross-validation model selection approach for choosing the regularization trade-off parameter(s) for each of the methods. In all our experiments, we select C over 5 hyper-parameter values $\{10^0, 10^1 \ldots, 10^4\}$ using 5×3 fold cross-validation. We set $C^\star$ to be 100 everywhere except in `SVM+`. We use a Gaussian RBF kernel for the MMD term with a fixed kernel width of 10.0. From what we observed, the `SVM+` baseline requires a broad range to infer its two hyper-parameters, C and $C^\star$, so we perform 5×3 fold joint cross-validation over the range $\{10^{-4}, 10^{-3}, \ldots, 10^4\}$.

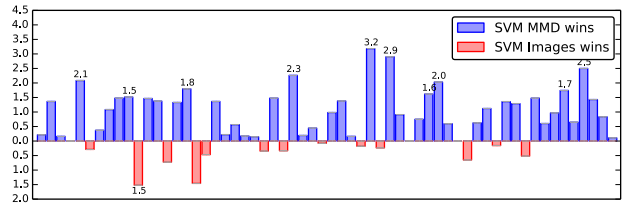
Evaluation metric. To evaluate the performance of the methods, we use the classification accuracy. We repeat each experiment 20 times using different random splits of the data into train and test sets and report mean and standard error across repeats.

5.1. Learning from the mistakes in abstract images

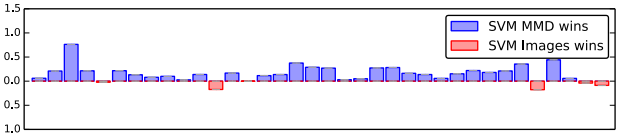
The INTERACT dataset contains 60 fine-grained classes that capture a variety of interactions between a pair of people, e.g., *running after*, *running to*, *arguing with*. Each of the interaction is represented as a set of real images and a set of clip art illustrations (on average, 50 images and 50 illustrations per class). The dataset has two settings: *category-level*, in which images and illustrations are collected independently given the category class, and *instance-level* where 2-3 illustrations are collected for a given image. Here, we detail the experimental results of the category-level setting, and the supplementary material contains the full table of results of the instance-level setting.

For each interaction class, we train a binary classifier to distinguish this interaction (positive class) against the remaining 59 interactions (negative class). To train a classifier, we randomly sample 25 positive vs 25 negative images, and for testing we use the remaining positive images balanced with the negative samples. For those methods that use privileged data (for training only), we take 50 clip art illustrations as positive samples (all available per class) and balance them with the clip art images from the negative

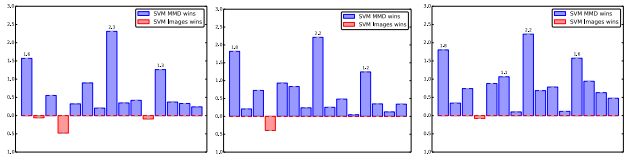
[1] https://computing.ece.vt.edu/~santol/projects/zsl_via_visual_abstraction/interact/index.html
[2] http://crcv.ucf.edu/data/UCF101.php
[3] http://geometry.cs.ucl.ac.uk/projects/2015/crosslink
[4] https://3dwarehouse.sketchup.com/?hl=en
[5] http://www.image-net.org



(a) Image + Clip art

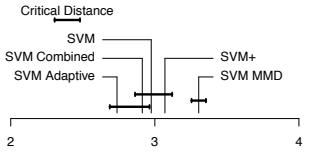


(b) Image + 3D model

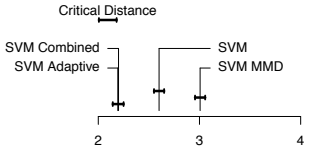


(c) Image+Video (spatial+motion) (d) Image+Video (spatial) (e) Image+Video (motion)

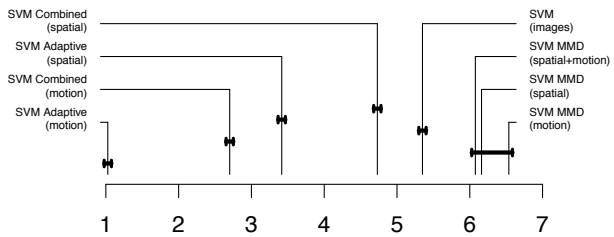
Figure 3: Learning *image* classifiers from the mistakes of classifiers trained on *other* datasets: *clip arts* (3a), *3D models* (3b), and *videos* (3c–3e) datasets. Pairwise comparison of the proposed SVM MMD and the baseline method (SVM trained on images). The height of the bar corresponds to the relative accuracy improvement over the baselines for each of the 60 one-versus-rest (Image + Clip art), 36 one-versus-one (Image + 3D model), and 15 one-versus-rest (Image + Video) problems. The full accuracy results are presented in Tables 1, 2, and in the supplementary material.



(a) Image + Clip art



(b) Image + 3D model



(c) Image + Video

Figure 4: Statistical summary of results based on Demšar [6]. Average rank of the methods (x axis) is computed based on accuracy (the higher the better). A critical distance measures significant differences between methods based on their ranks. We link two methods with a solid line if they are not statistically different (p-value $> 5\%$).

classes. In this experiment, to train SVM+, we randomly pair images and illustrations of the same class label to define the constraints in Eq. (1b). In the instance-level setting (in the supplementary), we use one corresponding illustration per image. We noticed that within an action category, the variability of clip art illustrations is moderate, and SVM+ performs similarly in category and instance-level settings.

In this dataset, real images and clip art illustrations are represented using 765 dimensional features that capture human poses, expressions, relation and appearance and are provided with the dataset. We use the features computed for Person A, who is performing the action with respect to Person B. In this case the privileged modality and the real images are expressed using the same feature representation, which makes it a perfect testbed to compare our proposed model with all baselines.

Results. The full result of this experiment is presented in Table 1 and the summary in terms of a pairwise comparison between the proposed SVM MMD and the standard SVM is in Figure 3a. We analyzed our experimental results on the INTERACT dataset using the multiple dataset statistical comparison method of [6] in Figure 4a. There is statistical evidence that SVM MMD performs best among the five methods. SVM+ performs better than SVM and SVM Combined in terms of ranking, however there is not enough evidence to support that the differences are significant. This conclusion holds true also for the instance-level setting (summarized in Figure 5). Advantageous performance of SVM MMD shows that learning from the mistakes of clip art classifiers help. We credit the significant improvement of SVM MMD over other methods to the principled idea of making the right decision at a larger scale by comparing error distributions rather than focusing on error in example-by-example basis. From the *average rank* perspective, SVM Combined and SVM Adaptive do not lead to an improved performance w.r.t. SVM Images. Looking closer at Table 1, there are cases when SVM Combined improves significantly (*sitting with*) and SVM Adaptive improves by a large margin (*elbowing*) but they are followed by large performance dips in cases such as *looking away from* (SVM Combined) and *talking with* (SVM Adaptive). In contrast, the drop in SVM MMD is only moderate at worst (action *waving at*).

5.2. Learning from the mistakes in videos

In this experiment, we cross video data from the UCF101 dataset with image data from the ImageNet dataset to address the task of action recognition in images. Both datasets intersect at sport activities, so we focus on the following 15 actions: *archery*, *basketball*, *biking*, *bowling*, *cricket shot*, *golf swing*, *horse riding*, *kayaking*, *pole vault*, *rafting*, *rowing*, *skateboard*, *skiing*, *surfing* and *tennis swing*. We collect images from the most relevant synsets in the ImageNet

	SVM Images	SVM Combined	SVM [39] Adaptive	SVM+ [35]	SVM (ours) MMD		SVM Images	SVM Combined	SVM [39] Adaptive	SVM+ [35]	SVM (ours) MMD
carrying	97.21 ± 0.33	94.50 ± 0.55	96.36 ± 0.46	$\mathbf{97.64 \pm 0.39}$	97.43 ± 0.46	crawling to	82.14 ± 0.79	79.02 ± 1.16	79.55 ± 1.19	$\mathbf{83.39 \pm 0.78}$	83.12 ± 0.76
catching	83.75 ± 1.15	84.20 ± 1.15	83.41 ± 1.18	84.20 ± 1.10	$\mathbf{85.11 \pm 1.10}$	jumping to	78.88 ± 1.15	$\mathbf{80.95 \pm 0.98}$	79.14 ± 1.06	80.00 ± 1.09	80.26 ± 1.26
pushing	80.08 ± 1.14	$\mathbf{82.74 \pm 0.90}$	81.37 ± 1.17	80.89 ± 0.93	80.24 ± 1.15	walking aw. fr.	78.87 ± 1.12	78.15 ± 0.80	76.69 ± 1.24	78.47 ± 0.86	$\mathbf{79.03 \pm 0.95}$
pulling	63.79 ± 0.89	$\mathbf{67.18 \pm 1.55}$	64.60 ± 1.42	62.82 ± 1.11	63.79 ± 1.30	running aw. fr.	84.29 ± 1.36	83.04 ± 0.77	84.64 ± 0.91	83.12 ± 1.21	84.11 ± 1.06
reaching for	66.42 ± 0.86	$\mathbf{68.58 \pm 1.02}$	67.17 ± 1.18	66.50 ± 1.36	68.50 ± 1.04	crawling aw. fr.	78.64 ± 1.48	80.23 ± 1.36	76.59 ± 1.84	77.73 ± 1.44	$\mathbf{81.82 \pm 1.60}$
jumping over	90.10 ± 0.95	93.17 ± 0.55	$\mathbf{93.46 \pm 0.84}$	90.77 ± 0.88	89.81 ± 0.86	jumping aw. fr.	$\mathbf{83.98 \pm 0.94}$	83.59 ± 0.87	82.66 ± 1.04	83.44 ± 1.00	83.75 ± 0.79
hitting	83.89 ± 1.39	83.70 ± 1.28	84.07 ± 1.24	83.43 ± 1.29	$\mathbf{84.26 \pm 1.11}$	walking after	84.20 ± 0.98	81.40 ± 0.98	82.80 ± 0.88	85.50 ± 1.11	$\mathbf{87.10 \pm 0.78}$
kicking	89.67 ± 1.05	$\mathbf{92.17 \pm 0.68}$	91.08 ± 0.81	90.75 ± 0.91	90.75 ± 0.79	running after	82.95 ± 1.03	81.44 ± 1.02	$\mathbf{84.17 \pm 0.98}$	83.79 ± 0.96	83.86 ± 0.71
elbowing	82.39 ± 1.00	84.43 ± 1.19	$\mathbf{86.82 \pm 0.87}$	83.86 ± 1.15	83.86 ± 0.95	crawling after	$\mathbf{86.67 \pm 0.91}$	84.88 ± 1.01	83.57 ± 1.04	86.19 ± 1.06	$\mathbf{86.67 \pm 1.04}$
tripping	86.36 ± 0.79	84.39 ± 0.96	85.23 ± 1.05	86.82 ± 0.61	$\mathbf{87.88 \pm 0.74}$	jumping after	81.42 ± 0.74	80.83 ± 0.78	79.00 ± 0.90	81.08 ± 0.85	$\mathbf{82.17 \pm 0.85}$
waving at	$\mathbf{71.20 \pm 1.29}$	67.72 ± 1.50	68.04 ± 1.43	70.33 ± 1.05	69.67 ± 1.11	walking past	80.15 ± 1.01	80.74 ± 0.81	78.53 ± 1.43	80.51 ± 0.91	$\mathbf{81.76 \pm 0.75}$
pointing at	77.33 ± 1.34	74.22 ± 1.47	73.79 ± 1.62	77.16 ± 1.38	$\mathbf{78.79 \pm 1.23}$	running past	76.09 ± 1.40	77.27 ± 1.18	75.23 ± 1.41	77.58 ± 1.45	$\mathbf{78.12 \pm 1.10}$
point. aw. fr.	66.50 ± 1.67	66.00 ± 1.74	63.62 ± 1.76	67.00 ± 1.39	$\mathbf{67.88 \pm 1.72}$	crawling past	78.10 ± 1.79	78.45 ± 1.29	77.62 ± 1.17	77.26 ± 1.65	$\mathbf{78.69 \pm 1.58}$
looking at	67.42 ± 1.37	$\mathbf{68.95 \pm 1.75}$	66.13 ± 1.28	67.50 ± 1.71	66.69 ± 1.49	jumping past	$\mathbf{77.41 \pm 1.31}$	73.15 ± 1.71	74.54 ± 1.70	76.67 ± 1.84	$\mathbf{77.41 \pm 1.45}$
looking aw. fr.	73.91 ± 1.61	67.66 ± 0.91	70.55 ± 1.25	73.12 ± 1.22	$\mathbf{75.23 \pm 1.29}$	stand. next to	84.35 ± 0.61	$\mathbf{84.89 \pm 0.71}$	81.63 ± 0.95	83.80 ± 0.78	83.70 ± 0.82
laughing at	72.34 ± 0.98	$\mathbf{74.22 \pm 1.28}$	71.95 ± 1.05	73.20 ± 0.95	74.14 ± 1.16	sitting next to	85.70 ± 1.03	84.69 ± 0.85	83.52 ± 0.83	84.77 ± 1.06	$\mathbf{86.33 \pm 0.90}$
laughing with	$\mathbf{82.29 \pm 0.90}$	81.88 ± 1.10	79.90 ± 1.32	80.21 ± 1.12	80.83 ± 1.05	lying next to	73.45 ± 1.32	73.88 ± 0.89	74.31 ± 1.16	73.36 ± 1.27	$\mathbf{74.57 \pm 1.10}$
hugging	87.89 ± 0.96	87.66 ± 1.16	88.36 ± 0.96	$\mathbf{88.44 \pm 0.75}$	87.42 ± 0.91	crouch. next to	80.31 ± 1.56	80.16 ± 1.00	78.59 ± 1.37	$\mathbf{80.78 \pm 1.55}$	80.16 ± 1.37
wrestling with	88.75 ± 0.87	85.80 ± 0.98	87.39 ± 0.81	89.66 ± 0.87	$\mathbf{90.11 \pm 0.95}$	stand. in fr. of	71.79 ± 1.46	67.14 ± 1.06	69.07 ± 1.28	72.07 ± 1.34	$\mathbf{73.14 \pm 1.18}$
dancing with	84.34 ± 0.82	81.10 ± 1.32	82.72 ± 1.09	83.82 ± 1.04	$\mathbf{84.56 \pm 0.82}$	sitting in fr. of	78.56 ± 0.95	79.70 ± 1.08	79.09 ± 0.96	78.48 ± 1.12	$\mathbf{79.85 \pm 1.02}$
hold. hands w.	85.48 ± 0.75	83.87 ± 0.99	84.60 ± 0.79	$\mathbf{86.13 \pm 0.71}$	86.05 ± 0.71	lying in fr. of	81.98 ± 1.01	81.12 ± 0.92	$\mathbf{83.10 \pm 0.86}$	82.76 ± 0.98	81.47 ± 1.16
shak. hands w.	94.74 ± 0.84	91.12 ± 0.69	92.76 ± 0.69	$\mathbf{95.69 \pm 0.51}$	94.91 ± 0.67	crouch. in fr. of	87.27 ± 0.91	87.05 ± 1.25	86.93 ± 1.20	86.70 ± 0.82	$\mathbf{88.75 \pm 0.78}$
talking with	82.06 ± 0.94	79.49 ± 1.39	76.54 ± 1.06	81.62 ± 1.40	$\mathbf{82.21 \pm 1.17}$	standing behind	71.03 ± 1.30	70.52 ± 1.55	$\mathbf{71.98 \pm 1.28}$	68.79 ± 1.47	71.64 ± 1.39
arguing with	83.62 ± 1.02	83.97 ± 0.80	$\mathbf{85.17 \pm 0.88}$	83.88 ± 1.13	83.28 ± 0.85	sitting behind	87.98 ± 0.79	$\mathbf{89.19 \pm 0.62}$	88.39 ± 0.79	88.63 ± 0.84	88.95 ± 0.81
walking with	91.93 ± 0.63	90.00 ± 0.89	89.43 ± 1.06	93.30 ± 0.57	$\mathbf{93.41 \pm 0.58}$	lying behind	80.68 ± 1.17	$\mathbf{83.11 \pm 1.09}$	82.27 ± 0.99	81.14 ± 1.08	82.42 ± 0.95
running with	$\mathbf{89.67 \pm 0.89}$	87.67 ± 1.22	89.58 ± 1.00	$\mathbf{89.67 \pm 0.88}$	89.33 ± 0.97	crouch. behind	76.30 ± 1.14	76.30 ± 1.11	75.09 ± 1.08	76.11 ± 0.84	$\mathbf{76.94 \pm 1.07}$
crawling with	79.76 ± 1.47	$\mathbf{82.38 \pm 0.85}$	80.12 ± 1.23	81.19 ± 1.46	82.02 ± 1.24	standing with	78.15 ± 1.39	78.79 ± 1.29	74.35 ± 1.48	79.03 ± 1.07	$\mathbf{80.65 \pm 0.96}$
jumping with	82.88 ± 0.96	$\mathbf{83.46 \pm 1.61}$	81.73 ± 1.56	81.92 ± 1.12	83.08 ± 0.83	sitting with	81.90 ± 1.21	$\mathbf{85.83 \pm 1.23}$	84.29 ± 1.17	82.50 ± 1.01	83.33 ± 1.06
walking to	78.75 ± 1.26	$\mathbf{79.73 \pm 0.81}$	77.68 ± 1.15	79.64 ± 1.04	79.20 ± 1.16	lying with	70.50 ± 1.13	70.92 ± 1.22	68.25 ± 1.24	$\mathbf{71.58 \pm 1.02}$	71.33 ± 1.14
running to	$\mathbf{78.12 \pm 1.10}$	77.81 ± 0.79	77.42 ± 0.97	77.81 ± 1.14	78.05 ± 1.20	crouch. with	79.78 ± 1.34	$\mathbf{81.09 \pm 1.01}$	80.33 ± 0.95	80.00 ± 1.08	79.89 ± 1.12
						avg. acc.	80.77	80.45	79.95	80.90	**81.49**

Table 1: Learning image classifiers with the mistakes of clip art classifiers (category-level setting). For instance-level setting, please refer to the supplementary material. The best result is highlighted in **boldface** with an extra blue for our SVM MMD.

	spatial+motion		**spatial**			**motion**		
	SVM Images	SVM (ours) MMD	SVM Combined	SVM [39] Adaptive	SVM (ours) MMD	SVM Combined	SVM [39] Adaptive	SVM (ours) MMD
Archery	83.87 ± 0.50	$\mathbf{85.44 \pm 0.30}$	83.48 ± 0.24	82.81 ± 0.36	$\mathbf{85.69 \pm 0.27}$	79.75 ± 0.70	73.47 ± 0.61	$\mathbf{85.67 \pm 0.28}$
Basketball	$\mathbf{91.95 \pm 0.33}$	91.89 ± 0.27	90.04 ± 0.35	87.60 ± 0.48	$\mathbf{92.16 \pm 0.26}$	89.82 ± 0.42	82.25 ± 0.49	$\mathbf{92.30 \pm 0.30}$
Biking	90.71 ± 0.24	$\mathbf{91.26 \pm 0.32}$	89.93 ± 0.33	88.49 ± 0.31	$\mathbf{91.44 \pm 0.20}$	86.87 ± 0.52	79.54 ± 0.54	$\mathbf{91.45 \pm 0.21}$
Bowling	$\mathbf{94.66 \pm 0.29}$	94.18 ± 0.38	93.24 ± 0.32	90.09 ± 0.47	94.27 ± 0.34	90.61 ± 0.59	85.72 ± 0.50	94.59 ± 0.34
CricketShot	84.96 ± 0.28	$\mathbf{85.29 \pm 0.66}$	83.77 ± 0.24	82.92 ± 0.33	$\mathbf{85.90 \pm 0.22}$	82.44 ± 0.43	78.97 ± 0.54	$\mathbf{85.84 \pm 0.22}$
GolfSwing	82.17 ± 0.49	$\mathbf{83.06 \pm 0.31}$	81.13 ± 0.41	80.03 ± 0.32	$\mathbf{83.00 \pm 0.30}$	80.51 ± 0.34	72.76 ± 0.53	$\mathbf{83.23 \pm 0.29}$
HorseRiding	90.39 ± 0.15	$\mathbf{90.60 \pm 0.18}$	90.18 ± 0.24	89.05 ± 0.32	$\mathbf{90.63 \pm 0.20}$	87.37 ± 0.55	79.56 ± 0.43	$\mathbf{90.49 \pm 0.25}$
Kayaking	83.14 ± 0.45	$\mathbf{85.45 \pm 0.19}$	82.27 ± 0.37	81.97 ± 0.46	$\mathbf{85.35 \pm 0.23}$	81.16 ± 0.47	75.28 ± 0.50	$\mathbf{85.37 \pm 0.26}$
PoleVault	87.86 ± 0.44	$\mathbf{88.20 \pm 0.41}$	84.41 ± 0.33	83.97 ± 0.37	$\mathbf{88.11 \pm 0.37}$	83.53 ± 0.66	77.34 ± 0.35	$\mathbf{88.54 \pm 0.39}$
Rafting	87.55 ± 0.29	$\mathbf{87.97 \pm 0.27}$	87.54 ± 0.27	86.35 ± 0.28	$\mathbf{88.03 \pm 0.16}$	85.40 ± 0.39	82.29 ± 0.38	$\mathbf{88.33 \pm 0.21}$
Rowing	$\mathbf{87.80 \pm 0.46}$	87.71 ± 0.36	$\mathbf{89.49 \pm 0.18}$	88.32 ± 0.23	87.84 ± 0.35	85.05 ± 0.52	80.13 ± 0.36	$\mathbf{87.91 \pm 0.31}$
SkateBoarding	81.13 ± 0.54	$\mathbf{82.39 \pm 0.42}$	81.46 ± 0.43	79.83 ± 0.48	$\mathbf{82.37 \pm 0.36}$	77.48 ± 0.67	72.32 ± 0.50	$\mathbf{82.70 \pm 0.40}$
Skiing	90.24 ± 0.45	$\mathbf{90.61 \pm 0.45}$	90.24 ± 0.24	88.49 ± 0.38	$\mathbf{90.59 \pm 0.31}$	86.93 ± 0.52	79.62 ± 0.39	$\mathbf{91.18 \pm 0.25}$
Surfing	88.52 ± 0.39	$\mathbf{88.85 \pm 0.28}$	88.64 ± 0.26	87.76 ± 0.32	$\mathbf{88.65 \pm 0.35}$	86.34 ± 0.26	82.87 ± 0.46	$\mathbf{89.15 \pm 0.21}$
TennisSwing	83.35 ± 0.47	$\mathbf{83.59 \pm 0.42}$	82.78 ± 0.32	81.62 ± 0.31	$\mathbf{83.69 \pm 0.36}$	79.49 ± 0.64	72.58 ± 0.67	$\mathbf{83.82 \pm 0.41}$
avg. acc.	87.22	**87.77**	86.57	85.29	**87.85**	84.18	78.31	**88.04**

Table 2: Learning image classifiers with the mistakes of video classifiers. Video data contain complementary information from still frames (spatial) and motion between frames (motion). The best result, for each of the video information (spatial, motion, and spatial+motion), is highlighted in **boldface** and an extra blue for our SVM MMD.

dataset to these action classes to form the image data modality. On average, each action class has 1000 images and 100 videos to train/test the models. Similarly to our previous experiment, we form 15 one-vs-rest binary classification tasks by randomly sampling 28 vs 28 images for training and 1000 vs 1000 images to test the methods (or as much as the class size allows). For those methods that use privileged data (for training only), we take all videos from the positive class and balance them with the same amount randomly sampled from the negative classes. As our image representation, we use 4096 dimensional features extracted from the fc7 activation layer in CaffeNet [17] fine-tuned on ImageNet VOC2012 [26]. As our video representation, we extract spatial and temporal representations from the Caffe models fine-tuned on ImageNet VOC2012 (the same as image data) and on optical flow of the UCF101 dataset as provided in [11]. This video representation allows us to study the effects of three types of privileged data in this scenario: the spatial signal alone (4096 dimensional), the motion signal alone (4096 dimensional), and the spatial+motion signals combined (8192 dimensional). The first two types, as they are in the same dimension as the image space, can be

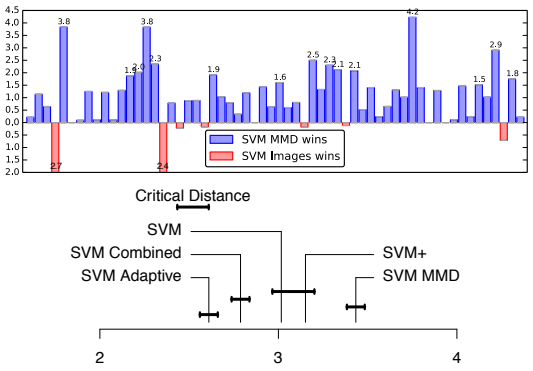


Figure 5: Learning image classifiers with the mistakes of clip art classifiers (instance-level setting). The full results are in the supplementary material.

used by the two domain adaptation baselines (SVM Adaptive and SVM Combined), whereas the spatial+motion information can not be used unless projection matrices are learned. Our SVM MMD does not depend on the dimensionality of the privileged data space.

Results. The results of this experiment are presented in Table 2, the summary in terms of a pairwise comparison between the proposed SVM MMD and the standard SVM is depicted in Figure 3c–3e, and statistical comparison of all methods is reported in Figure 4c. Overall, SVM MMD clearly improves over SVM Images in all settings of video information: spatial, motion, and spatial+motion. Specifically, in all cases but one, *bowling*, we can see positive improvements when using video modality as privileged information. The largest improvement appears when SVM MMD learns only from the mistakes of video classifiers with *motion* features. This can be credited to the complementary view of motion features w.r.t. the original image space and a good motion feature representation (deep features fine-tuned on optical flow of the UCF101 dataset).

5.3. Learning from the mistakes in 3D models

In contrast to our main task of object/action recognition in images, the CrossLink dataset was primarily designed to improve the performance of the 3D retrieval by leveraging images from the Bing search. We explore the setting where 3D models from the 3DWarehouse collection are used as privileged data to the images from the more complex ImageNet dataset. We collect 3D models by crawling the 3DWarehouse as described in [16], and manually checked all the models. Each 3D model is retrieved as a collection of 36 views of the object taken against no background. We use ImageNet synsets as our main image data. We focus on the following 9 classes: *airplane*, *backpack*, *bicycle*, *boat*, *car*, *chair*, *couch*, *helicopter*, *laptop*. Each object class has 1300 images on average and 90 3D models, ranging from 15 to 153 instances per class. As our image representation, we use 4096 dimensional deep features from the fc7 activation layer in CaffeNet fine-tuned on ImageNet VOC2012. As 3D model representation, we extract the same 4096 dimensional deep features from each of the views, and consider them as 36 data samples in the privileged space. For each pair of the 9 classes (36 in total) we train a one-vs-one binary classifier using 50 images (class balanced) for training and 2000 images (class balanced) for testing the models. For those methods that use privileged data, we balance 25 vs 25 instances of 3D models randomly sampled from the positive and negative classes.

Results. The full result of this experiment is presented in Table 1 of the supplementary material and the summary in terms of a pairwise comparison between the proposed SVM MMD and the standard SVM is in Figure 3b. Overall, SVM MMD improves over SVM Images (also supported by statistical summary in Figure 4b), but the actual improvement is minor. We credit this to the fact, that this recognition problem is rather simple, and the SVM Images baseline alone has achieved an average accuracy of 95.78%.

6. Conclusion and future work

> *A fool learns from his mistakes, but a truly wise man learns from the mistakes of others.*
>
> Otto von Bismarck

Learning with privileged information (LUPI) aims to exploit extra information that is available for *each instance* at training time. A typical assumption made is that these extra data are useful to discriminate between *easy* and *difficult* instances. We generalize this idea by describing a model that uses a divergence between distribution of *our own errors* and of *others' errors* as the loss function. Our approach can handle setting with no strict one-to-one correspondence between privileged and original data. We have shown the usefulness of this correspondence-free LUPI in the setting of cross-dataset learning of image classifiers. Our results reveal that learning image classifiers with the mistakes of clip art classifiers, or 3D classifiers, or video classifiers can be more accurate than learning using images only.

We seek to generalize our findings on correspondence-free LUPI for regression and multiple privileged information settings. We also aim to unify the LUPI setting and the setting where the extra attributes are available at test time but not at training time [18] under our framework of divergence minimization between the classifier errors in the privileged and original spaces. Finally, in the direction of deep model compression or distillation, we will assess the benefits of error matching as an alternative to matching the output class-probabilities commonly used in the literature.

Acknowledgment

We gratefully acknowledge NVIDIA for GPU donation and Amazon for AWS Cloud Credits.

References

[1] S. Antol, C. L. Zitnick, and D. Parikh. Zero-shot learning via visual abstraction. In *ECCV*, 2014. 3, 5

[2] J. Ba and R. Caruana. Do deep nets really need to be deep? In *NIPS*, 2014. 2

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, pages 2399–2434, 2006. 3

[4] J. Chen, X. Liu, and S. Lyu. Boosting with side information. In *ACCV*, 2013. 3

[5] K. Chwialkowski et al. Fast two-sample testing with analytic representations of probability measures. *NIPS*, 2015. 5

[6] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, pages 1–30, 2006. 6

[7] L. Duan, Y. Xu, W. Li, L. Chen, D. W. K. Wong, T. Y. Wong, and J. Liu. Incorporating privileged genetic info. for fundus image based glaucoma detection. In *MICCAI*, 2014. 1

[8] J. Feyereisl and U. Aickelin. Privileged information for data clustering. *Information Sciences*, pages 4–23, 2012. 1

[9] J. Feyereisl, S. Kwak, J. Son, and B. Han. Object localization based on structural svm using privileged information. In *NIPS*, 2014. 1

[10] S. Fouad, P. Tino, S. Raychaudhury, and P. Schneider. Incorporating privileged information through metric learning. *T-NNLS*, 2013. 1

[11] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, 2015. 7

[12] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-prob. In *NIPS*, 2006. 2, 4, 5

[13] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 2

[14] D. Hernández-Lobato, V. Sharmanska, K. Kersting, C. H. Lampert, and N. Quadrianto. Mind the nuisance: Gaussian process classification using privileged noise. In *NIPS*, 2014. 1, 2, 3

[15] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[16] M. Hueting, M. Ovsjanikov, and N. Mitra. Crosslink: Joint understanding of image and 3D model collections through shape and camera pose variations. *SIGGRAPH Asia*, 2015. 5, 8

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 7

[18] S. Khamis and C. H. Lampert. CoConut: Co-classification with output space regularization. In *BMVC*, 2014. 8

[19] W. Kienzle and K. Chellapilla. Personalized handwriting recognition via biased regularization. In *ICML*, 2006. 5

[20] M. Lapin, M. Hein, and B. Schiele. Learning using privileged information: SVM+ and weighted SVM. *Neural Networks*, pages 95–108, 2014. 1

[21] W. Li, L. Duan, D. Xu, and I. W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *T-PAMI*, pages 1134–1148, 2014. 2

[22] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016. 2

[23] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *IJCAI*, 2009. 2

[24] D. Pechyony and V. Vapnik. Fast optimization algorithms for solving SVM+. In *Stat. Learning and Data Science*, 2011. 1

[25] N. Quadrianto, J. Petterson, and A. Smola. Distribution matching for transduction. In *NIPS*, 2009. 2

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 1–42, 2015. 5, 7

[27] V. Sharmanska, D. Hernández-Lobato, J. M. Hernández-Lobato, and N. Quadrianto. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *CVPR*, 2016. 1

[28] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to rank using privileged information. In *ICCV*, 2013. 1, 2, 3, 4

[29] V. Sindhwani and D. S. Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *ICML*, 2008. 3

[30] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[31] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *JMLR*, pages 2389–2410, 2011. 4

[32] T. Tommasi and T. Tuytelaars. A testbed for cross-dataset analysis. In *ECCV Workshop*, 2014. 2

[33] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 2

[34] V. Vapnik and R. Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *JMLR*, pages 2023–2049, 2015. 1

[35] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, pages 544–557, 2009. 1, 2, 3, 4, 5, 7

[36] Z. Wang, X. Wang, and Q. Ji. Learning with hidden information. In *ICPR*, 2014. 3

[37] D. X. Wen Li, Li Niu. Exploiting privileged information from web data for image categorization. In *ECCV*, 2014. 1

[38] H. Yang and I. Patras. Privileged information-based conditional regression forest for facial feature detection. In *Automatic Face and Gesture Recognition (FG)*, 2013. 1

[39] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *ACM MM*, 2007. 5, 7

[40] X. Zhang, F. X. Yu, S.-F. Chang, and S. Wang. Deep transfer network: Unsupervised domain adaptation. *arXiv preprint arXiv:1503.00591*, 2015. 2