One-Shot Learning of Object Categories using Dependent Gaussian Processes

Erik Rodner and Joachim Denzler

Chair for Computer Vision
Friedrich Schiller University of Jena
{Erik.Rodner,Joachim.Denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. Knowledge transfer from related object categories is a key concept to allow learning with few training examples. We present how to use dependent Gaussian processes for transferring knowledge from a related category in a non-parametric Bayesian way. Our method is able to select this category automatically using efficient model selection techniques. We show how to optionally incorporate semantic similarities obtained from the hierarchical lexical database WordNet [1] into the selection process. The framework is applied to image categorization tasks using state-of-the-art image-based kernel functions. A large scale evaluation shows the benefits of our approach compared to independent learning and a SVM based approach.

1 Introduction

Learning an object category with a single example image seems to be a difficult task for a machine learning algorithm, but an easy everyday task for the human visual recognition system. A common hypothesis to justify the ability of the human cognition system to generalize quickly from few training examples is our use of prior knowledge from previously learned object categories [2]. This concept is known as *interclass* or *knowledge* transfer. In general, machine learning problems with few training examples are often highly ill-posed. Knowledge transfer from related categories allows to use prior knowledge automatically, which can be utilized to regularize such problems or enrich the training data set indirectly.

In the following we concentrate on knowledge transfer between binary classification tasks, which is also termed one-shot learning for the special case of a single training example. Given a target task with few positive training examples, one tries to select a support classification task from a heterogenous set of tasks with each having a relatively large number of training examples. These additional examples are then used to transfer prior knowledge to the target task.

Knowledge transfer techniques for image categorization were introduced by Fei-Fei et al. [3], who model knowledge as a prior distribution of the parameters of an object part constellation model. This prior distribution is used in a maximum-a-posteriori estimation of the target task model parameters. Tommasi and Caputo [4] present an extension to least-squares SVM which allows to adapt

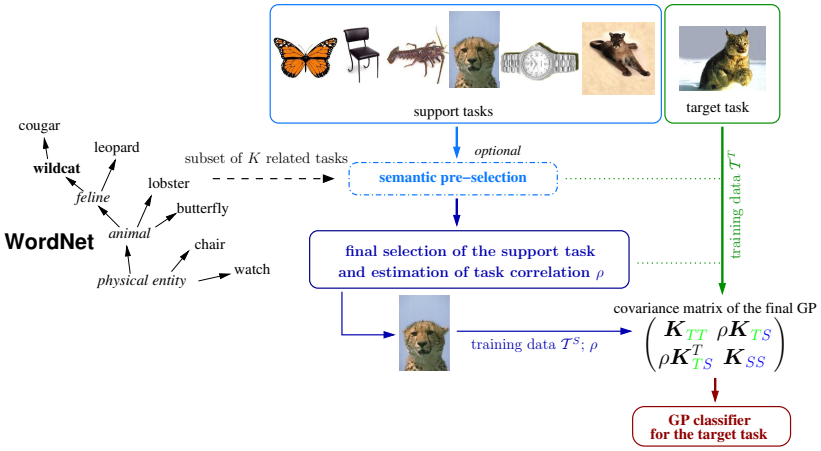


Fig. 1. Basic outline of the proposed transfer learning approach: Semantic similarities between categories and leave-one-out estimates are utilized to select a support task, which is used to transfer knowledge to a target task with dependent Gaussian processes.

the SVM solution of the target task to the decision boundary of a related object category. Our approach is based on classification and regression with Gaussian processes (GP), which has recently developed to a widely applied and studied machine learning technique [5] and is also used for image categorization [6]. One of the first papers investigating knowledge transfer with GP is the work of Lawrence et al. [7]. They show that the joint optimization of hyper-parameters using all tasks can be highly beneficial. Urtasun et al. [8] assume a shared latent feature space across tasks which can be also modeled in a GP framework.

We use dependent Gaussian process priors, as studied in [9, 10] and show how to utilize them for image categorization. Dependent GP priors allow us to efficiently transfer the information contained in the training data of a support classification task in a non-parametric way by using a combined (kernel) covariance matrix. The amount of information transferred is controlled by a single parameter estimated automatically which allows to move gradually from independent to complete combined learning. Parallel to our work, Cao et al. [11] used the same framework for machine learning problems, such as WiFi localization.

Additionally we handle the case of heterogenous tasks, where the set of available support tasks also includes unrelated categories, which do not contain any valuable information for the target task. Similar to [4], we utilize efficient leave-one-out estimates available for GP regression to select a single support classification task. We also show how to use similarities estimated with WordNet [1] to improve this selection. The basic steps of our approach are illustrated in Fig. 1.

The remainder of the paper is organized as follows. We will briefly review classification and regression with Gaussian processes, which is followed by describing transfer learning with dependent Gaussian processes. The question how to choose a valuable support task is answered in Sect. 3.1. Our choice of image-

based kernel functions is presented in Sect. 4. Experiments in Sect. 5 show the benefits of our approach in an image categorization application. A summary of our findings and a discussion of future research directions conclude the paper.

2 Classification with Gaussian Process Priors

In the following we will briefly review Gaussian process regression and classification. Due to the lack of space, we concentrate only on the main model assumptions and the resulting prediction equation. For a presentation of the full Bayesian treatment we refer to Rasmussen and Williams [5].

Given training examples $\boldsymbol{x}_i \in \mathcal{T}$, which denote feature vectors or images, and corresponding labels $y_i \in \{-1, 1\}$ we would like to predict the label y_* of an unseen example $\boldsymbol{x}_*$. The two main assumptions of Gaussian processes for regression or classification are:

1. There is an underlying latent function $\boldsymbol{f}$, so that labels y_i are conditionally independent given $\boldsymbol{f}(\boldsymbol{x}_i)$ and described using the so called noise model $p(y_i \mid \boldsymbol{f}(\boldsymbol{x}_i))$.
2. The function $\boldsymbol{f}$ is a sample of a Gaussian process (GP) prior $\boldsymbol{f} \sim \mathcal{GP}(\mathbf{0}, \mathcal{K}(\cdot, \cdot))$ with zero mean and covariance or kernel function $\mathcal{K}$.

The Gaussian process prior enables us e.g. to model the covariance of outputs $\boldsymbol{f}(\boldsymbol{x})$ as a function of inputs $\boldsymbol{x}$. With $\mathcal{K}$ being a kernel function describing the similarity of two inputs, one can model the common smoothness assumption that similar inputs should lead to similar function values and thus similar labels. The noise model can be quite general, and for classification tasks one often uses cumulative Gaussian or sigmoid functions [5]. In contrast, we will follow Kapoor et al. [6] and use a Gaussian noise model with variance σ^2

$$p(y_i \mid f(\boldsymbol{x}_i)) = \mathcal{N}(y_i \mid f(\boldsymbol{x}_i), \sigma^2) \tag{1}$$

which is the standard model for GP regression. The advantage is that we do not have to rely on approximated inference methods, such as Laplace approximation or Expectation Propagation. As we will show in Sect. 3.1, this also allows us to compute efficient leave-one-out estimates, which can be used for model selection. The treatment of the classification problem as a regression problem, which regards y_i as real-valued function values instead of discrete labels, can be seen as a clear disadvantage. Nevertheless, as shown in Nickisch et al. [12] the performance of this method is often comparable with Laplace approximation for classification tasks and is computationally much more efficient.

The GP regression model assumptions lead to analytical solutions for the prediction of the label y_*. Let $\boldsymbol{K}$ be the kernel matrix with pairwise kernel values of the training examples $\boldsymbol{K}_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{k}_*$ be kernel values $(\boldsymbol{k}_*)_i = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_*)$ corresponding to a test example $\boldsymbol{x}_*$. The GP model for regression leads to the following equation for the prediction $\bar{y}_*$:

$$\bar{y}_*(\boldsymbol{x}_*) = \boldsymbol{k}_*^T (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y} \ . \tag{2}$$

3 Transfer Learning with Dependent Gaussian Processes

We now consider the case that two binary classification tasks are given: a so called support task with a large amount of training data $\mathcal{T}^S$ and a target task with only few training examples $\mathcal{T}^T$. This setting is different from the scenario of multi-task learning in which one wants to train classifiers for multiple binary classification tasks in combination. In our case, we do not want to improve the classifier for the support task. This scenario for knowledge transfer is known as the concept of domain adaptation [4] or one-shot learning [3].

Our use of dependent Gaussian processes for transfer learning is based on the model proposed by Chai et al. [9]. For each task j we now have a latent function $\boldsymbol{f}_j$ which is assumed to be sampled from a GP prior. The key idea is that these functions are not assumed to be independent samples which allows us to transfer knowledge between latent functions. Thus, we use a combined latent function $\boldsymbol{f}((j,\boldsymbol{x})) = \boldsymbol{f}_j(\boldsymbol{x})$ which is a single sample of a GP prior with a suitable kernel function modeled by:

$$\mathcal{K}((j,\boldsymbol{x}),(j',\boldsymbol{x}')) = \begin{cases} \mathcal{K}^x(\boldsymbol{x},\boldsymbol{x}') & j = j' \\ \rho\, \mathcal{K}^x(\boldsymbol{x},\boldsymbol{x}') & j \neq j' \end{cases} , \tag{3}$$

with $\mathcal{K}^x$ being a base kernel function measuring the similarities of input examples. The hyper-parameter ρ of the extended kernel function with $0 \leq \rho \leq 1$ controls the correlation of the tasks: $\rho = 0$ corresponds to the case of independent learning whereas $\rho = 1$ assumes that the tasks are highly related. It should be noted that this type of knowledge transfer can also be motivated theoretically with a decomposition of the latent function into an average latent function shared by all tasks and an independent latent function [13].

We use only one single support classification task which is automatically selected using the techniques described in Sect. 3.1. In comparison to the single task GP model in equation (2), only the kernel function changes. Therefore, the label prediction of an unseen example $\boldsymbol{x}_*$ can be calculated as follows:

$$\begin{aligned} \bar{y}_*(\boldsymbol{x}_*) &= \boldsymbol{k}_*^T (\boldsymbol{K}(\rho) + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y} \\ &= \begin{pmatrix} \boldsymbol{k}_{T*} \\ \rho \boldsymbol{k}_{S*} \end{pmatrix}^T \left(\begin{pmatrix} \boldsymbol{K}_{TT} & \rho \boldsymbol{K}_{TS} \\ \rho \boldsymbol{K}_{TS}^T & \boldsymbol{K}_{SS} \end{pmatrix} + \sigma^2 \boldsymbol{I} \right)^{-1} \begin{pmatrix} \boldsymbol{y}_T \\ \boldsymbol{y}_S \end{pmatrix} , \end{aligned} \tag{4}$$

with $\boldsymbol{y}_{\mathcal{T}}$ and $\boldsymbol{y}_{\mathcal{S}}$ denoting the binary labels for the target and the support task respectively. The matrix $\boldsymbol{K}_{\mathcal{TS}}$ contains the pairwise kernel values of the target task and the support task. The same type of notational convention is used for $\boldsymbol{K}_{\mathcal{SS}}, \boldsymbol{K}_{\mathcal{TT}}$, $\boldsymbol{k}_{\mathcal{S}*}$ and $\boldsymbol{k}_{\mathcal{T}*}$.

Shared Background Category In the context of image categorization, one often has one single background and multiple object categories [14]. Thus, binary classification tasks share the background category. In this case $\mathcal{T}^S$ and $\mathcal{T}^T$ are not disjoint, which leads to an ill-conditioned kernel matrix $\boldsymbol{K}(\rho)$. We solve this problem by restricting the support training set only to examples of the object

category. Therefore the label vector $\boldsymbol{y}_{\mathcal{S}}$ is always a vector of ones. Please note that due to our zero mean assumption of the GP prior this leads to a valid classifier model and for the case of independent learning ($\rho = 0$) to an one-class GP classifier for the support task.

3.1 Selection of a Support Task using Leave-One-Out Estimates

The optimization of the hyper-parameter ρ and the selection of an appropriate support task can be handled as a combined model selection problem. To solve this problem, we use leave-one-out estimates similar to [4]. In the context of Gaussian process regression, the posterior of the label of a training example $\boldsymbol{x}_i$ conditioned on all other training examples can be computed in closed form [5]

$$\log p(y \mid \left(\mathcal{T}^T \cup \mathcal{T}^S\right) \setminus \{\boldsymbol{x}_i\}, \boldsymbol{y}, \rho) = -\frac{1}{2}\log \eta_i^2 - \frac{(y-\mu_i)^2}{2\eta_i^2} - \frac{1}{2}\log 2\pi \ , \quad (5)$$

with η_i^2 being the variance of the leave-one-out estimate μ_i:

$$\eta_i^2 = 1/\left(\boldsymbol{K}(\rho)^{-1}\right)_{ii} \quad \text{and} \quad \mu_i = y_i - \left(\boldsymbol{K}(\rho)^{-1}\boldsymbol{y}\right)_i \eta_i^2 \ . \quad (6)$$

The estimates μ_i offer to use a wide range of model selection criteria, such as leave-one-out log predictive probability [5] or squashed and weighted variants [4]. A common measure to assess the performance of a binary classification task is average precision [15]. Therefore, we use the calculation of the average precision directly using the estimates μ_i and ground truth labels y_i. This decision is additionally justified by experiments in the last paragraph of Sect. 5.2, which compares average precision to multiple model selection criteria embedded in our approach. Those experiments will also show that the conditional likelihood $p(\boldsymbol{y}^{\mathcal{T}} \mid \boldsymbol{y}^{\mathcal{S}}, \mathcal{T}^S, \mathcal{T}^T)$ is a non-appropriate model selection criterion in our setting.

We optimize the average precision with respect to ρ, which is a simple one-dimensional optimization, with golden section search [16] for each task of the set of given support tasks. The task and corresponding ρ value which yield the best average precision are chosen to build the final classifier according to equation (4).

3.2 Automatic Pre-Selection using WordNet

Selecting a support classification task among a large set of available tasks using only a single example is itself a very difficult problem, and the selection method described above, might not be able to transfer beneficial information. A solution is the use of prior knowledge from other information sources to pre-select tasks which are likely to be related.

We optionally use WordNet, which is a hierarchical lexical database of the English language, and the textual label of each object category. The usefulness of this information source has been demonstrated recently in the context of attribute based knowledge transfer [17] and hierarchical classification [18]. A possible assumption would be that semantically related object categories are also

visual similar. Thus the support task could be selected by semantic similarity measures such as the Reznik measure [1]. Whereas this assumption might hold for e.g. animal hierarchies, it might not hold in all cases and prevents important knowledge transfer from only visual similar tasks. Therefore we use WordNet in advance to leave-one-out selection and pre-select the K most related tasks among all available tasks based on their semantic similarity. For $K = 1$, WordNet selects the support task using the semantic of the category name and the leave-one-out method only optimizes ρ. If K equals the number of available support tasks, WordNet pre-selection does not influence transfer learning and the selection is based on visual similarity only. The importance of the combination of visual and semantic similarities for the selection will be analyzed empirically in Sect. 5.2.

4 Categorization Using Image-Based Kernels

One of the state-of-the-art feature extraction approaches for image categorization is the bag-of-features (BoF) idea. A quantization of local features which is often called a codebook, is computed at the time of training. We use OpponentSIFT [15] descriptors calculated on a dense grid and the method of Moosmann et al. [19] as the clustering method. For each image a histogram is calculated which counts for each codebook entry the number of matching local features. A standard way to apply the BoF idea to kernel-based classifiers is to use the calculated histograms as feature vectors and apply a traditional kernel function such as the radial basis function kernel.

In contrast, we define the kernel function directly on images. The spatial pyramid matching kernel as proposed by Lazebnik et al. [20] extends the BoF idea and divides the image recursively into cells (e.g. 2×2). In each cell the BoF histogram is calculated and the kernel value is computed using a weighted combination of histogram intersection kernels corresponding to each cell. In addition we use the gray-value based PHoG (pyramid histogram of oriented gradients) kernel [21] to compare our results directly to [4] in Sect. 5.1.

5 Experiments

Experiments are performed using all 101 object categories of Caltech 101 and a subset of the Caltech 256 database [3]. Both databases contain a large number of challenging object categories and a suitable background category. In each experiment a target task and corresponding few training images are selected. Training and testing is done for each target task 100 times with a random split of the data, which yields mean performance measure values. In our experiments we empirically support the following *hypotheses*:

1. Our transfer learning approach improves the mean performance compared to independent learning even with a large heterogenous set of available support classification tasks.
2. By using WordNet pre-selection, one can achieve a performance gain for nearly all classification tasks.

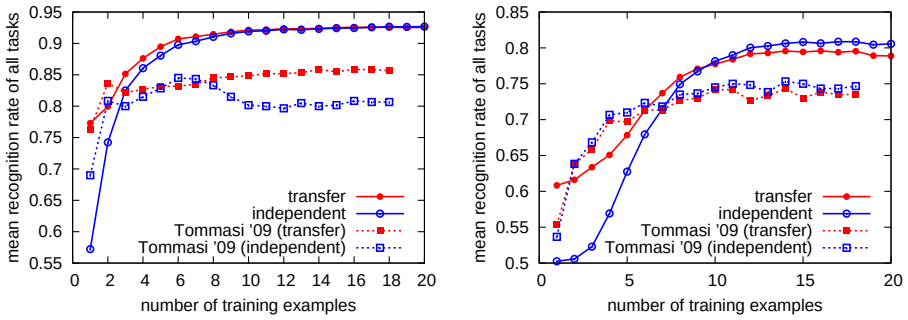


Fig. 2. Caltech 256 results for our transfer learning approach, independent learning and Adapted LS-SVM [4] (which performs a manual selection of adequate object images in advance): (Left) using *related* classification tasks, (Right) using *unrelated* tasks.

3. With a given set of related classification tasks, our method achieves higher recognition rates than Adapted LS-SVM [4].
4. Using the average precision of leave-one-out estimates as described in Sect. 3.1 yields the best performance among several other model selection criteria.

In contrast to multi-task learning with a shared training set [10], a non-zero noise variance σ is not essential to transfer knowledge. For this reason, we choose the noise variance σ^2 adaptively. We iteratively increase the value of σ^2 $(0, 10^{-8}, 10^{-7}, 10^{-6}, \ldots)$ until the Cholesky decomposition of the kernel matrix can be calculated ensuring its positive-definiteness.

5.1 Experiments with Caltech 256

We compare our approach to Adapted LS-SVM as proposed by [4] and tried to use an equivalent experimental setting. Two sets of classification tasks are chosen to study the cases of transferring knowledge using only related support classification tasks (car, fire-truck and motorbike) and using a heterogenous set of classification tasks (school-bus, dog and duck). Training and testing is done with a variable number of training images for the target object category and 18 training images for the background and support categories. It is important to note that in contrast to [4] we did not perform a manual selection of images, where the object is clearly visible without occlusion. To compare our results to [4] (values were extracted from Fig 1(a) and Fig 2(a) in the paper) we used the mean recognition rate of all tasks as a performance measure. A pre-selection of classification tasks using WordNet is not applied in this experiment.

Evaluation The results are shown in Fig. 2. First of all, it is clearly visible that learning benefits from knowledge transfer using our approach even in the "unrelated case" (*Hypothesis 1, page 6*). The same plots also validate that we are able to improve the results of [4] in the "related case" even by using images with occluded objects and different view points (*Hypothesis 3*). In the "unrelated case"

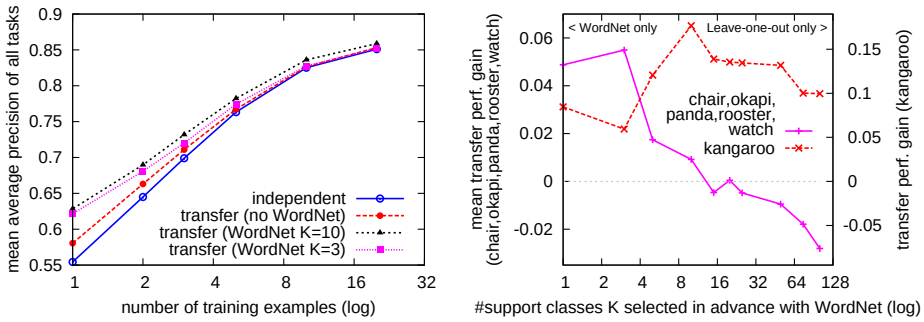


Fig. 3. (Left) Mean average precision of all tasks with a varying number of training examples. (Right) Different peaks of the one-shot learning performance for a varying number of support classes K pre-selected using WordNet: Mean average precision of tasks, which did not benefit from knowledge transfer without WordNet, and performance values of the kangaroo task for different values of K. The results of the kangaroo task highlights the importance of the combination of WordNet and our model-selection.

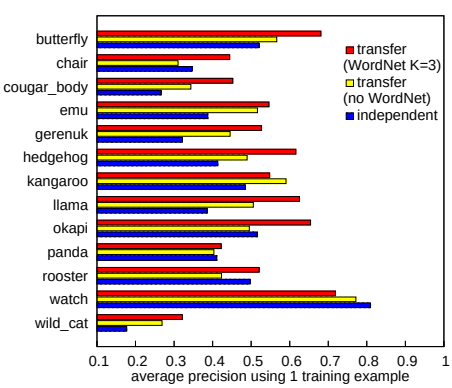
this unconstrained setting yields to lower results for independent learning, which makes transfer learning more important and leads to a significant performance gain even by using unrelated tasks. Our approach also improves [4] for the case of one-shot learning and when more than 7 training images are used.

5.2 Experiments with Caltech 101

In these experiments we use all 101 object categories as available support tasks and a subset of possible target tasks (listed in Fig. 4). As a performance measure for each binary classification task we use average precision as used in the Pascal VOC challenges [15]. Training and testing is done with a variable number of training images for the target object category, 30 training images for the support object categories and 200 background images.

Evaluation As to be seen in the left plot of Fig. 3 our transfer learning approach without WordNet pre-selection improves the mean average precision compared to independent learning when using few training examples and converges to it with more than 10 training examples (*Hypothesis 1*).

The detailed results for each task using a single training example are included in the left plot of Fig. 4 and deliver additional insight into the methods behavior: Transfer learning improves the average precision for some tasks significantly, e.g. task "gerenuk" with a performance gain of more than 11%, but also fails for some tasks like "okapi". This is due to a wrong selection of the support task using leave-one-out estimates and can be handled in almost all cases by using the WordNet pre-selection method (*Hypothesis 2*). Our transfer learning method fails for the task "watch", because there seems to be no related task in general. The right plot in Fig. 3 shows the benefit of WordNet for those cases by varying the number K of pre-selected support tasks. The same plot also highlights that



Are the left results significant compared to independent learning?
The p-value of the paired t-Test is below ... (*n.s.*= not significant):

	butterfly	chair	cougar_body	emu	gerenuk
WordNet K=3	1e-15	1e-15	1e-15	1e-15	1e-15
no WordNet	1e-7	1e-3	1e-8	1e-13	1e-12
	hedgehog	kangaroo	llama	okapi	panda
WordNet K=3	1e-15	1e-6	1e-15	1e-12	*n.s.*
no WordNet	1e-7	1e-11	1e-12	*n.s.*	*n.s.*
	rooster	watch	wild_cat		
WordNet K=3	*n.s.*	1e-13	1e-15		
no WordNet	1e-5	1e-4	1e-13		

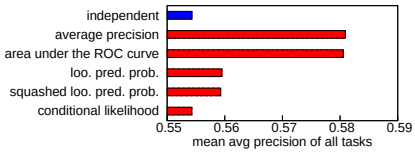


Fig. 4. (Left) Caltech 101 results for our transfer learning approach with and without pre-selection of support classification tasks using WordNet and for independent learning using a single training example. (Right) Mean average precision of all tasks using a single training example without pre-selection and different model selection criteria.

severe pre-filtering with WordNet ($K < 10$) leads to worse results for the task "kangaroo". The same holds for the mean average precision of all tasks which is lower for a strict pre-selection (with $K = 3$) compared to a pre-selection of only 10 support tasks (cf. left plot of Fig. 3). Therefore, only a combination of WordNet pre-selection with a selection based on leave-one-out estimates is reasonable when confronted with a new task.

We additionally evaluated our approach with different model selection criteria: average precision and area under the ROC curve using leave-one-out estimates, leave-one-out predictive probability [5] with squashed variants [4] and the conditional likelihood of the target task training set [11]. The results are shown in the right plot of Fig. 4, justifying our choice of average precision using leave-one-out estimates (*Hypothesis 4*).

6 Conclusions and Further Work

We presented an approach to transfer learning using dependent Gaussian processes, which is able to significantly improve the classification performance of one-shot learning and learning with few examples. Dependent Gaussian processes allowed us to express transfer learning in terms of a combined latent function with a suitable kernel function. Our method chooses a highly related classification task automatically by using the average precision achieved by leave-one-out estimates. We also studied the influence of the number of available tasks on the performance of the selection and demonstrated that an optional pre-selection of tasks using semantic similarities obtained from WordNet can be beneficial.

Further research has to be done to develop a more efficient model selection method to robustly estimate multiple hyper-parameters of the combined covariance function. For example, a combination of performance measures based on

leave-one-out estimates and standard maximum likelihood estimation might be suitable. Additionally, dependent Gaussian processes can also be used in conjunction with approximation methods for GP classification rather than regression.

References

1. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet: : Similarity - measuring the relatedness of concepts. In: Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2004). (2004) 38–41
2. Ahn, W., Brewer, W.F., Mooney, R.J.: Schema acquisition from a single example. J. of Experim. Psychology: Learning, Memory, and Cognition **18** (1992) 391–412
3. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. PAMI **28**(4) (2006) 594–611
4. Tommasi, T., Caputo, B.: The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In: BMVC. (2009)
5. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005)
6. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. IJCV **88**(2) (2009) 169–188
7. Lawrence, N.D., Platt, J.C., Jordan, M.I.: Extensions of the informative vector machine. In: Deterministic and Statist. Methods in Machine Learn. (2004) 56–87
8. Urtasun, R., Quattoni, A., Lawrence, N.D., Darrell, T.: Transfering nonlinear representations using gaussian processes with a shared latent space. In: Proceedings of the Learning Workshop (Snowbird). (2008) MIT-CSAIL-TR-2008-020.
9. Chai, K.M.: Generalization errors and learning curves for regression with multi-task gaussian processes. In: NIPS. (2009) 279–287
10. Bonilla, E., Chai, K.M., Williams, C.: Multi-task gaussian process prediction. In: NIPS, MIT Press (2008) 153–160
11. Cao, B., Pan, S.J., Zhang, Y., Yeung, D.Y., Yang, Q.: Adaptive transfer learning. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence. (2010)
12. Nickisch, H., Rasmussen, C.E.: Approximations for binary gaussian process classification. Journal of Machine Learning Research **9** (10 2008) 2035–2078
13. Pillonetto, G., Dinuzzo, F., Nicolao, G.D.: Bayesian online multitask learning of gaussian processes. PAMI **2** (February 2010) 193–205
14. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. PAMI **29**(5) (2007) 854–869
15. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI (in press) (2010)
16. Kiefer, J.: Sequential minimax search for a maximum. Proceedings of the American Mathematical Society **4**(3) (1953) 502–506
17. Rohrbach, M., Stark, M., Szarvas, G., Schiele, B., Gurevych, I.: What helps where - and why? semantic relatedness for knowledge transfer. In: CVPR. (2010)
18. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR. (2007) 1–7
19. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: NIPS. (2006) 985–992
20. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178
21. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: CIVR: Conference on Image and Video Retrieval. (2007) 401–408