Selecting Influential Examples: Active Learning with Expected Model Output Changes

Alexander Freytag*, Erik Rodner*, and Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Germany
{firstname.lastname}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. In this paper, we introduce a new general strategy for active learning. The key idea of our approach is to measure the expected change of model outputs, a concept that generalizes previous methods based on expected model change and incorporates the underlying data distribution. For each example of an unlabeled set, the expected change of model predictions is calculated and marginalized over the unknown label. This results in a score for each unlabeled example that can be used for active learning with a broad range of models and learning algorithms. In particular, we show how to derive very efficient active learning methods for Gaussian process regression, which implement this general strategy, and link them to previous methods. We analyze our algorithms and compare them to a broad range of previous active learning strategies in experiments showing that they outperform state-of-the-art on well-established benchmark datasets in the area of visual object recognition.

Keywords: active learning, Gaussian processes, visual recognition, exploration-exploitation trade-off.

1 Introduction

Over the last decade, the amount of accessible data has been growing dramatically and our community discovered the benefits of "big data" for learning robust recognition models [2,11]. However, in several important applications, *e.g.*, defect detection [15] or fine-grained categorization of rare categories [7], collecting labeled samples turns out to be a hard and expensive task, where labeling uninformative samples should be avoided as much as possible. Therefore, actively selecting an informative set of samples to label is important, especially if the labeling budget is strictly limited. Furthermore, it is necessary for *life-long learning* of visual objects, where we are interested in incrementally enriching object models with minimal user interaction.

In active learning, we are interested in reducing expensive manual labeling efforts. This goal is achieved by identifying a subset of unlabeled samples from a huge pool, such that the resulting accuracy (classification or regression accuracy depending on the application) of a model learned using the additional subset is maximized. The challenges are that the labels of the selected examples are only available after the selection

* A. Freytag and E. Rodner were supported by a FIT scholarship from the DAAD.

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 562–577, 2014.

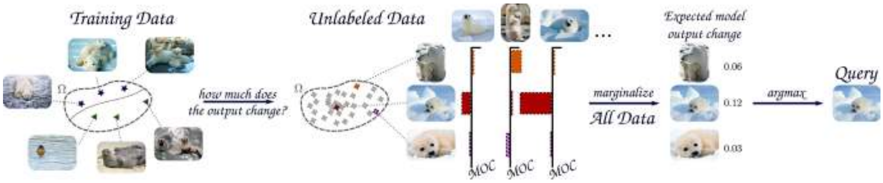

Fig. 1. Illustration of the active learning strategy introduced in this paper: an initial model is trained from labeled samples (blue and green). Unlabeled samples (gray) are evaluated with respect to the change of model outputs after adding them to the train set. For three exemplary samples (red, orange, and pink), the resulting model output change (MOC) for three different images is visualized. The sample leading to the strongest output change marginalized over all data Ω is finally queried.

and that the accuracy for unseen data can not be correctly measured beforehand and a proxy for it needs to be optimized. Although approximations exist based on estimated labels [16] or estimated confidence [1], they have to perform time-consuming extensive model evaluations and updates. Furthermore, estimating class labels is especially prone to errors in the presence of only few labeled data – which is the working range of nearly all active learning scenarios.

To circumvent these drawbacks, a variety of different strategies has been proposed, which are based on *what one assumes to be important* for higher accuracies, *e.g.*, a rapid exploration of the whole feature space [10], the identification of 'hard samples' among unlabeled points with respect to the current model [17], or combinations of existing techniques [1,4]. Although being intuitive in their different ways, none of these strategies can actually guarantee an impact of active learning on future model decisions.

In this paper, we therefore introduce a new general active learning strategy facing the problem by predicting the influence of an unlabeled example on future model decisions. If the unlabeled example is likely to change future decisions of the model when being labeled, it is regarded as an informative sample. An illustration of our strategy is visualized in Fig. 1. In the toy example, the final query is likely to change model outputs for a whole set of samples and is therefore preferred over samples which would lead to almost no changes. In summary, the contributions of this paper are two-fold: (1) we present a novel active learning strategy applicable to different applications and models, and (2) we derive an efficient algorithm based on Gaussian process regression from our general strategy.

The active learning strategy most similar to ours is to calculate the expected model change [6,20], which is mostly realized by measuring the Euclidean distance between the current model parameters and the expected model parameters after labeling. As we show in this paper, this strategy completely ignores the underlying data distribution, which is not the case for our approach, where we consider the change of model *outputs* instead of model parameters. Our technique is rather general and can be used for several different learning methods; however, we show that in the case of Gaussian process regression (or kernel ridge regression), the expected model output change can be efficiently calculated without learning from scratch. Furthermore, several active learning methods derived from the new approach are empirically compared with each other

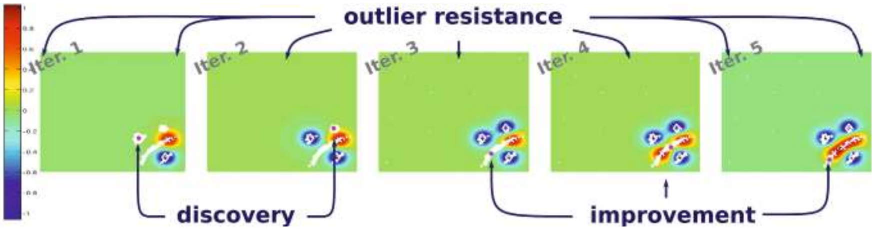

Fig. 2. Visualizing our active learning technique (EMOC). Figure is best viewed in color and by zooming in. Current classification scores are color coded, and thickness of unlabeled samples corresponds to their active learning scores, with large scores being preferred. You are invited to compare also against results of different strategies given in the supplementary material.

and we demonstrate in our experimental evaluation on well-known visual recognition datasets the advantage of expected model output change as a tool for active learning.

Why Yet Another Active Learning Technique? While facing active learning and diving through the enormous amount of great work done in this field, we sought for gaining clarity about which technique to use in which scenario. We investigated several playgrounds, among it a 2D toy example simple on first sight (see Fig. 2 and supplementary material). To our surprise, the majority of existing approaches struggled heavily due to either a focus too strong on outliers, or poor discovery abilities (see Table **1** in the suppl. mat. for an overview of our findings). The only positive exception, however, performed surprisingly poor in experiments on real-world data. As a consequence, we found the necessity to develop a technique inheriting advantages of previous active learning approaches, *i.e.*, being capable of discovering new clusters of data while being resistant to outliers, and on the same time focusing on regions where unconfident classification boundaries badly need improvements. In the remainder of the paper, we derive our technique from a theoretical position, prove it on real-world data, and, without further anticipating, show here already on the 2D problem that our technique offers the desired properties (see Fig. 2).

After reviewing related work in Sect. 2, we derive the main principle of our approach in Sect. 3. How to obtain efficient query strategies by utilizing Gaussian process models is shown in Sect. 4 and fast approximations are derived. Sect. 5 shows that the typical trick of density weighting in active learning can be motivated as a very rough approximation of our approach. We finally analyze the derived strategies on well-established benchmark datasets for visual object recognition in Sect. 6.

2 Related Work

As mentioned earlier, active learning techniques aim for selecting samples that lead to the highest improvement in accuracies, or similarly reduce the error as fast as possible. Since accuracies can not be reliably estimated in the absence of test data, multitudes of proxies and heuristics have been developed, which can be grouped into several general strategies. We review a prominent subset of them and list a few representative works.

Rapid Exploration. In order to quickly obtain label information for the whole feature space, exploration strategies prefer samples maximally far away from all current training samples, like KFF by [1] or the Gaussian process predictive variance [10]. However, since recognition problems are often characterized by low-dimensional manifolds of the original input space, these techniques often struggle with querying outliers rarely related to the class distribution.

Maximum Uncertainty. Similar to exploration, techniques relying on classifier uncertainty aim for selecting samples the current model is most uncertain about. In contrast to the previous strategy, this is done in a supervised manner taking the current model boundaries into account. Exemplary techniques are given in [21] and [10] for Support Vector Machines (SVM) or Gaussian process classifiers (GP), respectively.

Maximization of Expected Model Change. To balance exploration and exploitation and additionally ensure that queried samples affect the current model, techniques in this area favor samples that result in the largest model change after retraining. Since the process of model retraining is costly, techniques had been restricted to parametric models where the model change can be traced back to the change of the gradient of the objective function [20]. In our earlier work [6], we extended this strategy to non-linear GP regression models by exploiting efficient closed-form updates. However, a theoretical connection to the goal of error reduction is missing and assuming that a large model change results in an acceptable change of predictions is not valid in general.

Reduction of Estimated Classification Error. To overcome the previous shortcomings, this strategy directly aims at reducing the unknown classification error of the current model under a specified loss function. The technique most famous here was introduced by [16], where the true conditional distribution is approximated with the prediction of the current model, which leads to an expected entropy minimization scheme. Although being closest to the goal of active learning, techniques of this strategy suffer from two drawbacks: (i) they often have to face the computational costs of model retraining for every unlabeled sample and additionally have to evaluate the error on all available data, and (ii) the estimation of unknown labels needed for error evaluations is crucial and prone to errors especially in the presence of only few training data.

Work Most Similar to our Approach. The active learning approach we introduce in this paper is located *in between* the general strategies of *Maximization of expected model change* and *Reduction of estimated classification error*. While the first one does not take the actual change of model decisions into account at all, the latter requires perfectly reliable estimates of class labels used for empirical risk estimation. Both drawbacks are less present in our strategy. Additionally, we show that the density re-weighting technique introduced by [19] can be derived as an approximation of our proposed method. Furthermore, [22] present a technique to actively pick unlabeled nodes in a CRF to improve semantic segmentation quality based on the amount of CRF nodes flipping their state, which is generalized in our approach as well. For the choice of a GP regression model, we show how to transfer the work in [6] to our approach, which thereby additionally exploits the density information available in unlabeled data.

3 Expected Model Output Change (EMOC)

In pool-based active learning, a set $\mathfrak{L} \in \Omega^n$ of n labeled examples with labels $\boldsymbol{y} \in \mathcal{Y}^n$ and a set $\mathfrak{U} \subset \Omega$ of unlabeled examples is given from a problem domain Ω and an algorithm should select an example $\boldsymbol{x}' \in \mathfrak{U}$ to be labeled by an annotator, *i.e.*, assigned an output value $y' \in \mathcal{Y}$. The selection aims at improving the accuracy of a model $f : \Omega \to \mathcal{Y}$ (*e.g.*, a classifier for $\mathcal{Y} = \{-1, 1\}$ or a regressor for $\mathcal{Y} = \mathbb{R}$) learned by the given training data. In this paper, we deal with selecting one example at a time also known as myopic active learning.

As reviewed in the last section, several quite different active learning strategies exist. However, what we ultimately look for are high accuracy models with less annotated data, *i.e.*, we should select unlabeled examples $\boldsymbol{x}'$ that lead to the maximum increase in accuracy when being labeled and used to improve the current predictor. Unfortunately, we can never precisely predict the change in accuracy after adding a sample to the labeled pool in advance[1]. Therefore, we would like to raise the question: *Given a pool of unlabeled samples – some of them changing your model* outputs *when being labeled and added, others do not change anything – which one would you query?*

In absence of further knowledge, we argue that examples that lead to high expected model output changes should be queried. Therefore, we consider *how strongly a new sample $\boldsymbol{x}'$ influences the model decisions marginalized over all possible inputs $\boldsymbol{x} \in \Omega$ and over its yet unknown output $y' \in \mathcal{Y}$*:

$$\Delta f(\boldsymbol{x}') = \mathbb{E}_{y' \in \mathcal{Y}} \, \mathbb{E}_{\boldsymbol{x} \in \Omega} \left(\mathcal{L}\left(f_{(\mathfrak{L}, \boldsymbol{y})}(\boldsymbol{x}), f_{([\mathfrak{L}, \boldsymbol{x}'], [\boldsymbol{y}, y'])}(\boldsymbol{x}) \right) \right) \quad , \tag{1}$$

where $f_{(\cdot,\cdot)}$ is a model trained from labeled data and $\mathcal{L}$ is a loss function measuring the difference between model outputs. In the following, we skip the dependency of $\mathfrak{L}$, $\boldsymbol{y}$, $\boldsymbol{x}'$ and y' on f in the notation and instead write $f(\boldsymbol{x})$ and $f'(\boldsymbol{x})$ for models before and after including $(\boldsymbol{x}', y')$ as additional training sample:

$$\Delta f(\boldsymbol{x}') = \int_{\mathcal{Y}} \left(\int_{\Omega} \mathcal{L}(f(\boldsymbol{x}), f'(\boldsymbol{x})) \, p\,(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right) p(y'|\boldsymbol{x}') \mathrm{d}y'.$$

The active learning algorithm we propose evaluates $\Delta f(\boldsymbol{x}')$ for all unlabeled examples $\boldsymbol{x}' \in \mathfrak{U}$ and selects the example with the maximum value. In the following, we motivate the usefulness of this strategy for active learning and derive a specific algorithm from it by defining probability estimators, loss functions, and model classes.

EMOC Is an Upper Bound for Loss Reduction. Using EMOC for active learning can be motivated as an upper bound for the expected loss reduction. The additional assumption we need is that the loss function $\mathcal{L}$ obeys the following triangle inequality for $a, b, c \in \mathcal{Y}$:

$$\mathcal{L}(a, b) \leq \mathcal{L}(a, c) + \mathcal{L}(c, b) \quad . \tag{2}$$

[1] Note that we can not even precisely measure the accuracy of our system before the update and instead have to rely on approximations based on validation and test sets.

Furthermore, we assume that the model $g(\boldsymbol{x})$ learned from all possible data of the problem domain Ω exists[2] and that

$$\epsilon_f = \mathbb{E}_{\boldsymbol{x} \in \Omega} \left(\mathcal{L} \left(f(\boldsymbol{x}), g(\boldsymbol{x}) \right) \right) \quad , \tag{3}$$

is giving us an error measure for f. The expected decrease in loss for f' can now be defined as $\Delta\epsilon = \mathbb{E}_{y' \in \mathcal{Y}}(\epsilon_f - \epsilon_{f'})$ and the following shows that the expected model output change defined in Eq. (1) is an upper bound [22]:

$$\begin{aligned} \Delta\epsilon &= \mathbb{E}_{y' \in \mathcal{Y}} \, \mathbb{E}_{\boldsymbol{x} \in \Omega} \left(\mathcal{L}(f(\boldsymbol{x}), g(\boldsymbol{x})) - \mathcal{L}(f'(\boldsymbol{x}), g(\boldsymbol{x})) \right) \\ &\leq \mathbb{E}_{y' \in \mathcal{Y}} \, \mathbb{E}_{\boldsymbol{x} \in \Omega} \left(\mathcal{L}(f(\boldsymbol{x}), f'(\boldsymbol{x})) \right) = \Delta f(\boldsymbol{x}'). \end{aligned}$$

It is impossible to directly maximize the loss reduction term on the left-hand side for active learning, because g is unknown. Therefore, our active learning methods search for the unlabeled example $\boldsymbol{x}'$ with highest upper bound in loss reduction given by EMOC. It should be noted that this of course does not guarantee a proper decrease in the loss, but it at least does not limit it in advance by selecting examples that do not change model outputs at all.

Possible Choices for $\mathcal{L}(\cdot, \cdot)$ and $p(\boldsymbol{x})$. The choice of $\mathcal{L}$ naturally complies with the problem settings faced. The absolute difference $|f(\boldsymbol{x}) - f'(\boldsymbol{x})|$ in model response is well suited for regression tasks with continuous output values. For classification tasks, where $\mathcal{Y}$ is a discrete set, the common choice for measuring model output changes would be the classification loss, where $\mathcal{L}(a, b)$ is 1 for $a \neq b$ and zero everywhere else. However, we will see that for classification decisions based on thresholding underlying continuous model outputs, simpler losses can be used that avoid estimating a threshold (Sect. S1 in the supplementary material). Marginalization over y' is done by computing estimates based on the current model output and we discuss the details in the next section. In the future, we are planning to investigate also losses suitable for multi-class classification tasks.

What remains is how to model the probability distribution over the input space in practice, *i.e.*, how to specify $p(\boldsymbol{x})$. Since we only have access to the set $\mathcal{S} = \mathfrak{L} \cup \mathfrak{U}$ of labeled examples $\mathfrak{L}$ and unlabeled examples $\mathfrak{U}$, we approximate $p(\boldsymbol{x})$ with the empirical density distribution:

$$p(\boldsymbol{x}) \approx \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_j \in \mathcal{S}} \delta(\boldsymbol{x} - \boldsymbol{x}_j) \quad , \tag{4}$$

where $\delta(\boldsymbol{x})$ is the Dirac function. Note that this implies a representative data distribution in $\mathcal{S}$, which however is one of the main assumptions for active learning. In summary, EMOC scores (see Eq. (1)) for models with continuous outputs can be calculated based on empirical estimates for given data by:

$$\Delta f(\boldsymbol{x}') = \mathbb{E}_{y' \in \mathcal{Y}} \left(\frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_j \in \mathcal{S}} |f(\boldsymbol{x}_j) - f'(\boldsymbol{x}_j)| \right) \quad , \tag{5}$$

independent of the learning algorithm for f.

[2] For classification, we need this requirement for g because the label space is not $\mathbb{R}$. However, for regression, we could even assume that g is the ground-truth function.

4 Efficient EMOC with GP Regression

Our active learning strategy introduced in the previous section applies to a broad span of possible models and a naive approach to calculate the scores is to train a new predictor f' for each unlabeled example $\boldsymbol{x}' \in \mathfrak{U}$ by adding the example to the training set, retrain the predictor, and evaluate it on the whole set of examples given. In the following, we show that this is not necessary when using Gaussian process regression models. Furthermore, we also show how the marginalization over y' can be directly done.

Gaussian process regression is a kernel approach with the following decision function:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \; \kappa(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\alpha}^T \boldsymbol{k}(\boldsymbol{x}) \tag{6}$$

where κ is a given kernel function. The weight vector $\boldsymbol{\alpha}$ of the model is the result of kernel ridge regression and given by $\left(\mathbf{K} + \sigma_{\mathsf{n}}^2 \cdot \mathbf{I}\right)^{-1} \boldsymbol{y}$, where $\mathbf{K}$ is the kernel matrix of the training set $\mathfrak{L}$, $\boldsymbol{y}$ is the vector of outputs of $\mathfrak{L}$, and σ_{n}^2 their assumed noise variance [13]. For the model in Eq. (6), the influence of a new sample on all possible predictions can therefore be computed as follows:

$$\begin{aligned} \Delta f(\boldsymbol{x}') &= \mathbb{E}_{y' \in \mathcal{Y}} \left(\frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_j \in \mathcal{S}} |\boldsymbol{k}(\boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{\alpha} - \bar{\boldsymbol{k}}(\boldsymbol{x}_j)^{\mathrm{T}} \bar{\boldsymbol{\alpha}}| \right) \\ &= \mathbb{E}_{y' \in \mathcal{Y}} \left(\frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_j \in \mathcal{S}} |\bar{\boldsymbol{k}}(\boldsymbol{x}_j)^{T} \left(\begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} - \bar{\boldsymbol{\alpha}} \right) | \right) \end{aligned}$$

where $\bar{\boldsymbol{k}}(\boldsymbol{x}) = [\boldsymbol{k}(\boldsymbol{x}) \quad \kappa(\boldsymbol{x}', \boldsymbol{x})]$ and $\bar{\boldsymbol{\alpha}} = \bar{\mathbf{K}}^{-1} \begin{bmatrix} \boldsymbol{y} \\ y' \end{bmatrix}$ is the updated weight vector computed using the regularized kernel matrix $\bar{\mathbf{K}}$ of $\mathfrak{L} \cup \{\boldsymbol{x}'\}$.

Efficient Model Updates. Instead of computing expected model output changes from scratch by retraining the model with each unlabeled example, GP regression allows us to compute the new model and therefore also the change $\Delta\boldsymbol{\alpha}$ of model coefficients efficiently and in closed form for a given output y':

$$\Delta\boldsymbol{\alpha} \doteq \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} - \bar{\boldsymbol{\alpha}} = \frac{y' - \boldsymbol{k}^{\mathrm{T}} \boldsymbol{\alpha}}{\sigma_{f_*}^2 + \sigma_{\mathsf{n}}^2} \begin{bmatrix} \left(\mathbf{K} + \sigma_{\mathsf{n}}^2 \cdot \mathbf{I}\right)^{-1} \boldsymbol{k} \\ -1 \end{bmatrix}, \tag{7}$$

where $\sigma_{f_*}^2$ is the predictive variance of $\boldsymbol{x}'$ and $\boldsymbol{k}$ is the vector of pairwise kernel values of the training set and $\boldsymbol{x}'$. A proof is given in [6] and is based on block-wise matrix inversion. Please note that similar derivations for linear models are also possible and known for several decades [12].

EMOC for Classification with Label Regression. With GP regression as an illustrative example, we focused on models with a continuous output. For classification, we assume the final output to be obtained by maximizing a given $p(y \mid f(\boldsymbol{x}))$. In fact, a prominent set of popular models used for binary classification, such as SVMs [18], GP-regression [13], or LDA [9], first compute continuous scores for test samples which are then transformed to discrete responses, *e.g.*, by comparing against an application-specific threshold. For the case of GP classification, using GP regression scores as proposed by [10] leads to treating outputs as continuous values with Gaussian random noise and allows for skipping approximate inference necessary for direct GP classification [13]. Since we focus on binary classification settings in the rest of this paper, we argue to compute model output changes directly on the continuous scores, which reflects the non-ordinal nature of these models, avoids tricky threshold determinations, and will also be important to develop a fast version of our active learning algorithms. Furthermore, we will see in the experiments that label regression leads to an even better classification performance than proper classification models with approximate inference (supplementary material, Sect. S1).

Although our computed model output changes are based on continuous scores, the labels y' are still binary, *i.e.*, $y' \in \{-1, 1\}$, and we need a method to compute the expectation with respect to y' in Eq. (1). We know that the output of GP regression is not only a deterministic estimate but rather a predictive Gaussian distribution with mean $f(\boldsymbol{x}')$ and variance $\sigma^2_{f_*}$. Therefore, we compute $p(y' = 1|\boldsymbol{x}')$ by calculating the probability that a sample from this distribution is positive, which directly corresponds to the manner in which classification decisions are done in [10][3]. This probability can be calculated in closed form:

$$p(y' = 1|\boldsymbol{x}') = \frac{1}{2} - \frac{1}{2} \cdot \mathrm{erf}\left(-f(\boldsymbol{x}')/\sqrt{2\sigma^2_{f_*}}\right) \tag{8}$$

using the error function erf(z) and is related to the cumulative Gaussian noise model presented by [13]. In summary, the inner part of $\Delta f(\boldsymbol{x}')$ is evaluated using Eq. (7) for $y' = 1$ as well as $y' = -1$, and both values are used to compute a weighted average using the probability in Eq. (8).

Fast Approximated EMOC. Even with the efficient model updates presented, we have to compute $\bar{\boldsymbol{k}}(\boldsymbol{x}_j)^T \Delta\boldsymbol{\alpha}$ for every unlabeled example $\boldsymbol{x}_j$, which can be a huge computational burden for large sets of unlabeled examples. Fortunately, the computation time necessary to evaluate EMOC can be significantly reduced by the following approximation:

$$\Delta f(\boldsymbol{x}') = \mathbb{E}_{y' \in \mathcal{Y}}\Big(\frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_j \in \mathcal{S}} |\bar{\boldsymbol{k}}(\boldsymbol{x}_j)^T \Delta\boldsymbol{\alpha}|\Big) \tag{9}$$

$$\leq \mathbb{E}_{y' \in \mathcal{Y}}\Big(\frac{1}{\mathcal{S}} \sum_{\boldsymbol{x}_j \in \mathcal{S}} \bar{\boldsymbol{k}}(\boldsymbol{x}_j)^T \mathrm{abs}(\Delta\boldsymbol{\alpha})\Big) \tag{10}$$

[3] A more sophisticated estimation might also marginalize over possible thresholds, and future work should focus on integrating this aspect efficiently.

where $\mathrm{abs}(\cdot)$ denotes element-wise absolute value. Since the model change $\Delta\boldsymbol{\alpha}$ itself is independent of $\boldsymbol{x}_j$, we can write this in short form as

$$\Delta f_{\text{fast}}(\boldsymbol{x}') = \mathbb{E}_{y'}\Big(\sum_{i=1}^{n+1}\Big(|\Delta\alpha_i|\,\frac{1}{\mathcal{S}}\sum_{\boldsymbol{x}_j\in\mathcal{S}}\kappa\left(\boldsymbol{x},\boldsymbol{x}_j\right)\Big)\Big). \tag{11}$$

The computational benefit of the approximation introduced above is two-fold: (i) the asymptotic runtime reduces (see Table 2), and (ii) instead of demanding the kernel matrix consisting of kernel values between every two samples of $\mathcal{S}$, computing the approximated version only needs the resulting row sums. Additionally, note that the approximation has an interesting interpretation: the right-hand side can be seen as a Parzen density estimation

$$\mathrm{PDE}(\boldsymbol{x};\mathcal{S}) \propto \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}_j\in\mathcal{S}}\kappa(\boldsymbol{x},\boldsymbol{x}_j) \tag{12}$$

of each of the corresponding training samples estimated with both labeled and unlabeled data ($\mathcal{S} = \{\mathfrak{L}\cup\mathfrak{U}\}$). Therefore, the changes of model coefficients are weighted with respect to the data likelihood, *i.e.*, an outlier with a high change in $\boldsymbol{\alpha}$ should not have a huge impact on EMOC and therefore also not on the selection process during active learning. Note that kernel density estimation takes place in a possibly high-dimensional input space, however, we have not seen any issues with respect to the curse of dimensionality in our experiments.

5 Density Weighting as a Special Case of EMOC

In the following section, we do not propose any novel method, but show that a further approximation of the EMOC principle leads to density-weighted queries [19]. This connection emphasizes the importance of density weighting for active learning in general, which will be further studied in our experiments. Density weighting has been known for quite a while for active learning, but to the best of our knowledge, we are the first ones presenting it as an approximation of a very general active learning strategy.

Let us now denote the vector containing density values of all labeled samples with $\boldsymbol{p}_{\mathfrak{L}} : \boldsymbol{p}_{\mathfrak{L}}^{(i)} = \mathrm{PDE}(\boldsymbol{x}_i;\mathcal{S})$ and let further be $p_{\boldsymbol{x}'} = \mathrm{PDE}(\boldsymbol{x}';\mathcal{S})$ the data density value of the new sample. We can then even further approximate $\Delta f_{\text{fast}}(\boldsymbol{x}')$:

$$\Delta f_{\text{fast}}(\boldsymbol{x}') = \mathbb{E}_{y'}\left(\left|\Delta\boldsymbol{\alpha}^{\mathrm{T}}\cdot\begin{bmatrix}\boldsymbol{p}_{\mathfrak{L}}\\ p_{\boldsymbol{x}'}\end{bmatrix}\right|\right) \tag{13}$$

$$\leq \mathbb{E}_{y'}\left(\|\Delta\boldsymbol{\alpha}\|_1\cdot\left\|\begin{bmatrix}\boldsymbol{p}_{\mathfrak{L}}\\ p_{\boldsymbol{x}'}\end{bmatrix}\right\|_1\right) \tag{14}$$

$$\propto \mathbb{E}_{y'}\Big(\|\Delta\boldsymbol{\alpha}\|_1\cdot|p_{\boldsymbol{x}'}|\Big)\ , \tag{15}$$

where we used the fact that only terms depending on $\boldsymbol{x}'$ are important for the selection process during active learning.

Consequently, we notice that this very rough and simplified approximation of our proposed active learning strategy is equivalent to taking queries based on expected model change, *e.g.*, using [6], and to multiply the scores with a Parzen density estimate of the corresponding unlabeled sample. This indeed seems intuitive, since we now can ensure that the samples being queried not only affect the model, but are also likely to occur in dense regions of the space with respect to our current subspace of interest. Note that in contrast to [6], we marginalize over y' instead of using only the most likely y' as proposed by the authors.

Extension to Arbitrary Query Strategies. Based on the previous ideas, we will use the following straight-forward replacements of arbitrary query strategies[4] $\mathcal{Q}(\boldsymbol{x})$ in order to integrate the data density:

$$\mathcal{Q}^*(\boldsymbol{x}) = p_{\boldsymbol{x}} \cdot \mathcal{Q}(\boldsymbol{x}) \ . \tag{16}$$

The suggested replacement is a heuristic, which is easy to apply and motivated by our approach. We thereby ensure that samples of high density areas[5] are preferred over outliers from non-important, sparse regions. We show in our experiments that this modification improves the performance of previous active learning methods and therefore also offers a fairer comparison to our new active learning methods based on expected model output change.

Note that the idea for density-based re-weighting of query scores was already introduced in [19] for the task of sequence labeling. However, the authors proposed it as a heuristic without a clear theoretical motivation and simply note that they would *recommend it in practice*. In contrast to that, we place this technique in a proper theoretical background by deriving it as a special approximation of the more general active learning approach introduced in this paper. Furthermore, the Parzen estimates in Eq. (12) use all the given samples $\mathcal{S} = \{\mathfrak{L} \cup \mathfrak{U}\}$ in contrast to $\mathcal{S} = \mathfrak{U}$ as proposed in [19]. This is reasonable, since (1) density estimates do not have to be adapted during the query process, which naturally changes $\mathfrak{U}$ and (2) we estimate densities for the actual problem setting, *i.e.*, we rely on all information and data we have and not only on an arbitrary subset.

6 Experimental Results

For an experimental evaluation, we have been interested in the following aspects (i) how do active learning techniques derived from our strategy compare to previous state-of-the-art strategies (see Sect. 6.1), (ii) how strongly does the density-based re-weighting affect learning accuracies of state-of-the-art techniques (see Sect. 6.2), (iii) how powerful is the EMOC approach under ideal settings, where we assume a perfect label estimation for model updates (see Sect. 6.3), and is label regression enough for classification (supplementary material, Sect. S1)? To answer these questions, we follow the experimental setup of [6] and use the corresponding evaluation protocol[6]. We conduct

[4] If $\mathcal{Q}$ is designed to query samples with minimum score, multiply by $1 - p(\boldsymbol{x})$ instead.

[5] Density is considered with respect to the current problem and its induced data distribution.

[6] Evaluation protocol was taken from `https://github.com/cvjena/activeLearning-GP`

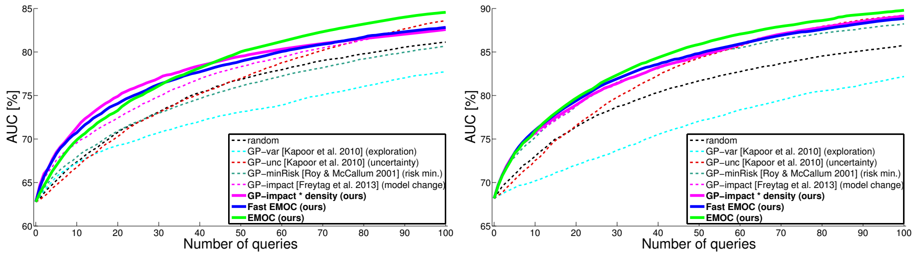

Fig. 3. Active Learning on binary tasks derived from **ImageNet** (*left*) and **Caltech 256** (*right*) . We compare against active learning strategies 'rapid exploration', 'maximum uncertainty', 'maximization of expected model change', and 'reduction of estimated classification error'. See text for further details. The figure is best viewed in color.

experiments on two well-established benchmark datasets for image categorization: ImageNet [3] and Caltech-256 [8]. Source code of our techniques and experiments will be made publicly available at our homepage `http://www.inf-cv.uni-jena.de/active_learning`.

6.1 Comparison to State-of-the-Art

Experimental Setup. As done in [6], we derive 100 random binary tasks from the ImageNet challenge consisting of a single positive and 9 negative classes. Every task is repeated with 10 random initializations, which finally results in 1,000 experiments to allow for reliable conclusions. Images are represented using the publicly available bag-of-words-features of ILSVRC 2010[7] and we evaluate our EMOC approach when using GP regression models as introduced in Sect. 4. In particular, we rely on the exact model output change as introduced in Eq. (4) (*EMOC*), its approximation given in Eq. (13) (*Fast EMOC*), as well as the density-weighted model change (*GP-impact* · *density*, see Eq. (15)), which is an extension of [6] derived from the EMOC principle. We compare against passive learning (*random*), which is the naive baseline for all active learning settings. Apart from that, we chose representative state-of-the-art methods for the reviewed strategies: *GP-var* (using examples with a high predictive GP variance) and *GP-unc* (seek for small ratios of predictive GP mean and variance) have been introduced by [10] and are representative for the rapid exploration and maximum uncertainty strategy, respectively. Additionally, we compare against the technique introduced in [6] (*GP-impact*) for the expected model change strategy, and the approach of [16] transferred to GP (*GP-minRisk*) for reduction of estimated classification error principle. See Sect. S4 in the supplementary material for further details of the experimental setup.

Evaluation. Active learning curves on tasks derived from ImageNet are shown in the left plot of Fig. 3. Dashed curves correspond to strategies existing in the literature, whereas our techniques are plotted in solid lines. First of all, we observe that the rapid exploration strategy leads to worse classification accuracies then passive learning,

[7] `http://www.image-net.org/challenges/LSVRC/2010`

Table 1. Experimental results for ImageNet and Caltech 256: Average AUC value (in %) after 50 queries for **100** random binary **tasks** averaged over 10 random initializations. $(^*)$ Our approach significantly outperforms all other approaches verified by a paired t-Test and $p < 10^{-3}$.

Strategy	ImageNet	Caltech 256
Random	76.83	81.70
Predictive variance [10] (GP-var)	73.11	77.06
Classification uncertainty [10] (GP-unc)	76.97	84.31
Reduction of classification error [16] (GP-minRisk)	76.08	84.59
Model change [6] (GP-impact)	78.32	84.56
EMOC strategy (Ours)*	**80.03**	**85.88**

which indicates the preference of outliers being queried. In contrast, nearly all remaining methods improve random sampling. Interestingly, empirical risk minimization leads to results slightly inferior to passive learning, emphasizing the negative influence of wrongly estimated labels. In contrast, queries based on expected model output change result in a significant improvement, confirming the intuition that samples resulting in different model *responses* are worth being labeled. As argued in Sect. 4, this partly originates from the fact that previous methods focus on less important aspects of unlabeled examples, *e.g.*, looking for unexplainable samples (as done by GP-uncertainty) might result in *interesting* samples, but without a clear relation to improvement of accuracy.

To further verify our findings, we performed experiments with an identical setup on the Caltech-256 dataset and visualized results in the right plot of Fig. 3. The results clearly indicate the before-mentioned relation between different active learning approaches. Uncertainty-based strategies obviously have problems querying useful samples if the size of training data is relatively small. Apart from that, the results are consistent with the observations from the ImageNet experiment. In addition, Table 1 contains average AUC values obtained after 50 queries. Our approach outperforms all other strategies significantly (paired t-Test, $p < 10^{-3}$). It has to be emphasized again that we obtained this result from experiments with 100 different recognition tasks and 10 different initializations. Furthermore, we compare against representative approaches of five active learning strategies.

With respect to the approximations of the EMOC principle, we further observe that for few labeled data in the ImageNet experiment, the approximation performs better then the exact EMOC scores, whereas for a larger number the relation switches. This suggests that the introduced approximations are less affected by randomly initialized training sets, whereas exact EMOC scores are especially valuable when the current model can be trained more robustly.

Evaluation in a One-vs-All Scenario. We also evaluated our approach in binary tasks created in a one-vs-all manner as done by [10] (see supplementary material for further details). As can be seen in Fig. 4, we observe a similar performance benefit of our methods as in the previous experiments. Visualizations of queried images for the airplane task as well as some hand-picked, interesting queries for remaining scenarios are given in the supplementary material (see Sect. S5, Fig. 1).

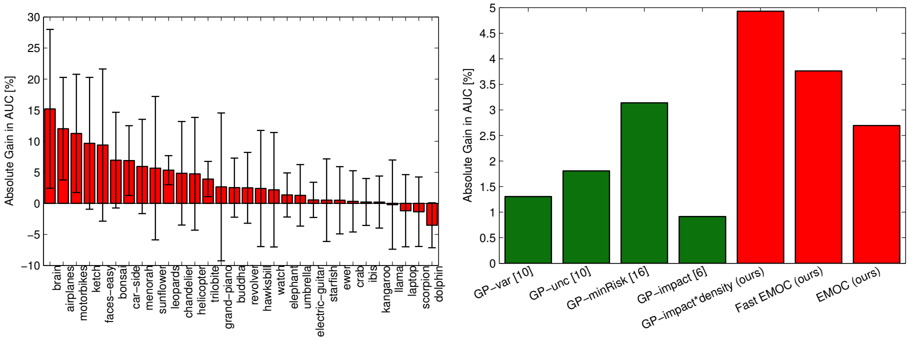

Fig. 4. Active learning improvement over passive baseline after 20 queries for 30 one-vs-all binary tasks derived from Caltech-256. *Left:* Results on all 30 individual classes with our Fast EMOC technique. *Right:* Averaged results for all compared techniques.

Table 2. Computation times needed performing active selection with our method and approximations (Sect. 4) and n labeled examples as well as u unlabeled examples

Method	**GP-impact** [6]	[6] · **density**	**Fast EMOC**	**EMOC**
Time ($n = 10, u = 990$)	$7.81 \cdot 10^{-4}$ s	$7.64 \cdot 10^{-4}$ s	$8.08 \cdot 10^{-4}$ s	$2.38 \cdot 10^{-2}$ s
Asymptotic time	$\mathcal{O}(n^2\mathbf{u})$	$\mathcal{O}(n^2\mathbf{u})$	$\mathcal{O}(n^2\mathbf{u})$	$\mathcal{O}(n^2 u + n\mathbf{u^2})$

Runtime Evaluation. An empirical comparison of computation times[8] for the derived methods when querying the first sample in the previous experiments is given in Table 2. As expected, exact EMOC computation is significantly slower than its approximated versions, which are as fast as existing strategies. Furthermore, the asymptotic time reveals that in case of a large number u of unlabeled examples, Fast EMOC should be the method of choice, because the selection time only depends linearly on the number of unlabeled samples. The approach of [16] has the same asymptotic time as exact EMOC, but we observed a speed-up of 1.6 over [16] in practice.

6.2 Importance of Density-Based Re-weighting

As mentioned earlier, density-based re-weighting of query scores is not limited to our introduced strategies, but can be extended to arbitrary active learning techniques. In the left plot of Fig. 5, we show the resulting gains when applying the re-weighting scheme to GP active learning strategies introduced in the literature. First of all, we observe that learning results are improved in almost all cases. Apart from this, it is also intuitive why some of the methods can only benefit from the heuristic in late stages of learning (like *GP-var*), whereas others can draw a partial advantage in early stages (*e.g.*, *GP-impact*). For example, as stated in [6], relying on the highest estimated model change leads to an implicit balancing between exploration and exploitation. Consequently, in early stages of learning, where the exploration aspect is usually more important, the density-based

[8] Computation times have been measured using a Matlab implementation on a 3.4 GHz CPU without parallelization and excluding precomputed kernel values.

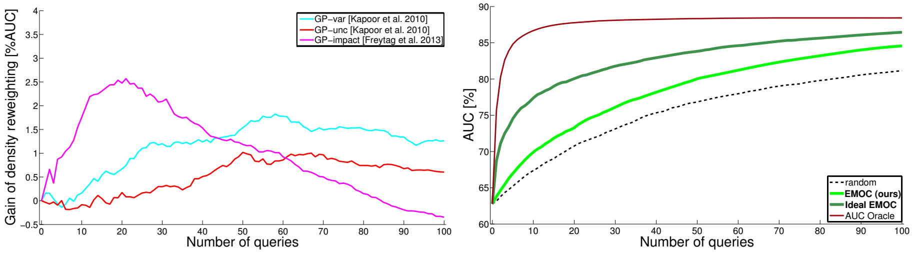

Fig. 5. *Left*: Gain of density-based re-weighting of active learning scores for several active learning strategies. *Right*: Performance of our EMOC method in a real active learning setting with unknown label y' (EMOC), when label y' is known during selection with EMOC (Ideal EMOC), and when the ground-truth label is used to greedily maximize AUC (AUC oracle).

re-weighting leads to exploration more focused on dense clusters than on outliers. In later stages, where exploration is less necessary, a focus too strong on dense regions is also less important and might even decrease performance.

6.3 Ideal EMOC

In a last experiment, we analyze the performance of our approach for a perfect estimation of model updates, *i.e.*, an artificial setup with known labels y'. Although we could directly use them for performance optimization on $\mathfrak{U}$, we are instead interested in the upper bound for EMOC. Therefore, we follow the previous experimental setup, and replace the expectation over possible labels in Eq. (1) by ground-truth labels for the unlabeled pool. As can be seen in the right plot of Fig. 5, working with correct label estimations again significantly improves learning rates. This observation is interesting by considering the experimental results in Sect. 6.1, where the EMOC strategy already outperformed existing techniques. Consequently, it would be highly beneficial to better infer unknown labels, especially if only few labeled samples are available and the initial model might be learned poorly. The right plot of Fig. 5 also contains the results of a perfect AUC oracle, where examples are chosen that maximize the AUC performance on $\mathfrak{U}$. The plot of this method is the upper bound for all myopic active learning methods.

7 Conclusions

We presented a new general active learning strategy based on calculating the expected change of model outputs when an unlabeled example would be labeled and incorporated. The main motivation is that examples with a high overall impact on the model *outputs* are most informative during learning. Our approach is flexible and allowed us to derive several new active learning methods. In particular, we showed how to compute expected model output changes efficiently for Gaussian process regression models. An extensive experimental evaluation revealed that our strategies outperform several existing active learning techniques on established benchmark datasets for image categorization. We further showed that density-based re-weighting of arbitrary active learning scores can be derived as a rough approximation of our introduced approach and presented its benefits. We conclude that our general strategy for active learning – to query samples which lead to the highest change of model responses – is beneficial in scenarios, where collecting labeled data is expensive or time-consuming.

For future research, several directions are possible: (1) combining our strategy with others in a reinforcement learning scheme [1,4], (2) improved estimation of class labels for unlabeled samples using semi-supervised learning [23], (3) active learning for regression and multi-class classification tasks by testing other loss functions, and (4) using sparsification techniques or efficient kernel evaluations [14,5] to speed up evaluation in the presence of large-scale data, such as the whole ImageNet dataset

References

1. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. Journal of Machine Learning Research (JMLR) 5, 255–291 (2004)
2. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 71–84. Springer, Heidelberg (2010)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition, CVPR (2009)
4. Ebert, S., Fritz, M., Schiele, B.: Ralf: A reinforced active learning formulation for object class recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3626–3633 (2012)
5. Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Rapid uncertainty computation with gaussian processes and histogram intersection kernels. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 511–524. Springer, Heidelberg (2013)
6. Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Labeling examples that matter: Relevance-based active learning with gaussian processes. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 282–291. Springer, Heidelberg (2013)
7. Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: Conference on Computer Vision and Pattern Recognition, CVPR (accepted for publication, 2014)
8. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694. California Institute of Technology (2007)
9. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classfication. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 459–472. Springer, Heidelberg (2012)
10. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. International Journal of Computer Vision (IJCV) 88, 169–188 (2010)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, NIPS (2012)
12. Plackett, R.L.: Some theorems in least squares. Biometrika 37(1/2), 149–157 (1950)
13. Rasmussen, C.E., Williams, C.K.I.: Adaptive Computation and Machine Learning. Adaptive Computation and Machine Learning. The MIT Press, Cambridge (2006)
14. Rodner, E., Freytag, A., Bodesheim, P., Denzler, J.: Large-scale gaussian process classification with flexible adaptive histogram kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 85–98. Springer, Heidelberg (2012)

15. Rodner, E., Wacker, E.S., Kemmler, M., Denzler, J.: One-class classification for anomaly detection in wire ropes with gaussian processes in a few lines of code. In: Conference on Machine Vision Applications (MVA), pp. 219–222 (2011)
16. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: International Conference on Machine Learning (ICML), pp. 441–448 (2001)
17. Schohn, G., Cohn, D.: Less is more: Active learning with support vector machines. In: International Conference on Machine Learning (ICML), pp. 839–846 (2000)
18. Schölkopf, B., Smola, A.J.: Learning with kernels: Support Vector Machines, Regularization, Optimization, and beyond. Adaptive Computation and Machine Learning. The MIT Press, Cambridge (2002)
19. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1070–1079. Association for Computational Linguistics (2008)
20. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Advances in Neural Information Processing Systems (NIPS), pp. 1289–1296. MIT Press (2008)
21. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research (JMLR) 2, 45–66 (2002)
22. Vezhnevets, A., Buhmann, J.M., Ferrari, V.: Active learning for semantic segmentation with expected change. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3162–3169 (2012)
23. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In: Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining (ICML-WS), pp. 58–65 (2003)