Large-scale Active Learning with Approximations of Expected Model Output Changes

Christoph Käding[1,2], Alexander Freytag[1,2], Erik Rodner[1,2], Andrea Perino[3,4], and Joachim Denzler[1,2,3]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Michael Stifel Center Jena, Germany
[3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
[4]Institute of Biology, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

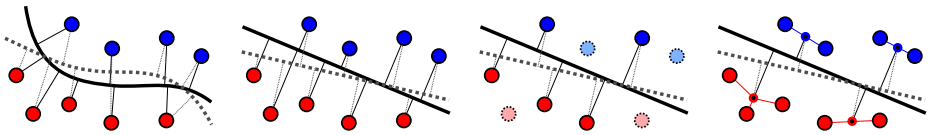
Abstract. Incremental learning of visual concepts is one step towards reaching human capabilities beyond closed-world assumptions. Besides recent progress, it remains one of the fundamental challenges in computer vision and machine learning. Along that path, techniques are needed which allow for actively selecting informative examples from a huge pool of unlabeled images to be annotated by application experts. Whereas a manifold of active learning techniques exists, they commonly suffer from one of two drawbacks: (i) either they do not work reliably on challenging real-world data or (ii) they are kernel-based and not scalable with the magnitudes of data current vision applications need to deal with. Therefore, we present an active learning and discovery approach which can deal with huge collections of unlabeled real-world data. Our approach is based on the expected model output change principle and overcomes previous scalability issues. We present experiments on the large-scale MS-COCO dataset and on a dataset provided by biodiversity researchers. Obtained results reveal that our technique clearly improves accuracy after just a few annotations. At the same time, it outperforms previous active learning approaches in academic and real-world scenarios.

1 Introduction

Over the past years, we observed striking performance leaps in supervised learning tasks due to the combination of linear models and deep learnable image representations. However, the demand for annotated data grew in the same frequency of newly published accuracy records. On the other hand, our ability to provide increasing labeled datasets is limited. Similarly intuitive is the observation that not all labeled images are equally informative for a given task. The area of active learning tackles this observation: by designing algorithms which estimate the gainable information of unseen examples, annotation costs can be reduced by labeling only the most informative ones.

While active learning has been an active area of research for more than 20 years [6], the majority of algorithms have been evaluated on synthetic or small scale datasets (*e.g.,* [6,27,16,18,14]). Undoubtedly, these algorithms achieve reasonable results in the

This research was supported by grant DE 735/10-1 of the German Research Foundation (DFG).



(a) Kernel EMOC [8] (b) Linear EMOC (c) Linear EMOC + Random Sampling (d) Linear EMOC + Clustering

Fig. 1: Visualization of the expected model output change (EMOC) criterion and our approximations thereof. Classification boundaries in a two-class scenario are marked as thick lines before (solid) and after a model update (dashed). Model outputs are denoted as thin lines. In [10], all output changes of a kernelized model are computed (a). A linear model reduces complexity of model evaluations (b). Evaluating output changes only on a random subset (c) or on cluster centroids (d) reduces computational complexity further.

presented benchmarks. However, it is unclear whether their performances with respect to computation time, memory demand, and classification accuracy scale sufficiently to large real-world datasets. However, especially due to the increasing availability of large unlabeled datasets, today's active learning algorithms have to have "large-scale abilities". In this paper, we follow this observation by presenting an active learning technique which is able to deal with large-scale datasets as well as unseen classes.

Our approach is based on the expected model output change (EMOC) criterion by Freytag et al. [10], which we transfer to linear models and approximate appropriately. Hence, the contribution of this paper is two-fold:

1. We present a realization of EMOC for regularized linear least square regression and derive efficient closed-form solutions for score computation and model update, and
2. we introduce and analyze approximations of resulting scores to reduce computational burdens even further.

The combination of both complementary aspects overcomes limitations of previous approaches and enables active learning with sophisticated selection criteria for large sets of unlabeled data. A schematic overview of our approaches is given in Fig. 1. We evaluate our techniques on the challenging MS-COCO dataset [21] and provide evidence for the technique's suitability in large-scale scenarios. Finally, we present results on a real world dataset from biodiversity researchers who are particularly interested in classifying large collections of unlabeled images.

2 Related Work

Active learning is a widely studied field which aims at reducing labeling efforts. The general goal of active learning is to optimally select examples for annotation, which ultimately reduces the error of the classifier as fast as possible. A detailed summary of active learning techniques was presented by Settles [26]. In the following, we only review the most prominent and most relevant techniques.

Established Approaches for Active Learning The early work of Roy and McCallum [24] presented approximations of estimated error reduction. Since the resulting algorithm is still computationally expensive, a majority of follow-up work has been based on surrogate functions, *e.g.,* [27,2,16,26,15,18,9,5]. A common strategy is the selection of examples the current classifier is most uncertain about [27,15,16,18]. This strategy is for example also used by [29,7] and further extended to query whole batches of unlabeled data. Complementary is the idea of rapidly exploring the space [2,18] which is purely data-driven. A third example is the preference of examples which leads to large changes of model parameters [9,26,5]. However, these concepts miss a clear connection to the reduction of errors, which is the ultimate goal of learning.

Related Work on Large-scale Active Learning The necessity for dealing with large datasets is a well-known problem for active learning. A simple yet efficient solution is to sample batches of data for evaluation, as presented by Hoi et al. [13] and Fu and Yang [11]. Similarly, the data can be pre-clustered [1]. Alternatively, models need to be applied which scale well to large datasets, *e.g.,* the probabilistic k-nearest neighbors approach by Jain et al. [15]. Another challenge is the growing number of classes. By only taking a subset of possible updates into account, Ertekin et al. [8] showed how this can be handled efficiently. Although these approaches scale nicely to large datasets, they still miss a clear connection to the reduction of errors.

Approaches using Expected Output Changes In contrast to the previous approaches, selecting examples which lead to large estimated model output changes approximates the expected reduction of errors. Motivated by Vezhnevets et al. for the task of semantic segmentation [28], it was later presented by Freytag et al. [10] from a general perspective. The authors initially focused on specific realizations for binary classification with Gaussian process models. Later on, they extended their approach to multi-class scenarios with unnameable instances [17]. Although the reported results lead to impressive accuracy gains in challenging scenarios, the authors were limited to medium-scale datasets due to the choice of kernel classifiers. In this work, we show how to overcome these limitations using two complementary aspects: (i) by transferring the EMOC criterion to linear models and (ii) by approximating involved expectation operations. The approach of [4] proposes to estimate an optimal mix based on expected model output changes of passive and arbitrary active learning techniques. The idea of their method is orthogonal to ours and can be used in addition to the techniques presented in our paper.

3 Expected Model Output Changes in a Nutshell

In this paper, we aim at extending the recently proposed expected model output changes (EMOC) criterion to large-scale scenarios. We start with a short review of the underlying idea and its multi-class variant as presented in [10,17].

The EMOC Principle In [10], Freytag et al. introduced EMOC as an approximation to the estimated reduction of expected risk. In contrast to previous approaches, the criterion favors only those examples $\mathbf{x}'$ which lead to largest output changes after re-training – averaged over any possible label $y' \in \mathcal{Y}$:

$$\Delta f\left(\mathbf{x}'\right) = \mathbb{E}_{y' \in \mathcal{Y}}\, \mathbb{E}_{\mathbf{x} \in \Omega}\left(\mathcal{L}\left(f\left(\mathbf{x}\right), f'\left(\mathbf{x}\right)\right)\right) \quad . \tag{1}$$

Here, f' refers to the updated model after including the unlabeled example $\mathbf{x}'$ with the estimated label y' to the current training set. The output changes are calculated with a loss function $\mathcal{L}$, *e.g.,* using an L_1-loss as suggested in [10]. However, note that Eq. (1) is not computable in practice for realistic settings due to the expectation over the input space Ω.

Similar to the transfer from expected risk to empirical risk, the authors of [10] proposed to rely on the empirical data distribution induced by the labeled data $\mathfrak{L}$ and unlabeled data $\mathfrak{U}$. For classification scenarios, we further note that labels are discrete. Combining both aspects leads to the final EMOC criterion for arbitrary classification scenarios with finite data:

$$\Delta f\left(\mathbf{x}'\right) = \sum_{y' \in \mathcal{Y}} \left(\frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x} \in \mathfrak{L} \cup \mathfrak{U}} \mathcal{L}\left(f\left(\mathbf{x}\right), f'\left(\mathbf{x}\right)\right) \right) \; p(y'|f\left(\mathbf{x}'\right)) \quad . \tag{2}$$

As it turned out, this active learning criterion is more robust compared to expected risk minimization and achieves state-of-the-art results.

EMOC for Multi-class Scenarios with Unnameable Instances Besides the theoretical derivation of EMOC, [10] contained efficient realizations only for Gaussian process models in binary classification tasks. Later on, the authors extended their results to multi-class scenarios by specifying appropriate loss functions and estimators for multi-class classification probabilities [17]. Furthermore, they investigated scenarios where unlabeled data contains so-called "unnameable instances". Unnameable instances refer to examples for which an oracle can not provide a proper label. These instances, also termed "noise" examples, will not lead to any improvement of the classifier and should thus be avoided during selection. The work of [17] shows that this can be achieved by (1) density re-weighting and (2) predicting unnameable instances by learning their distribution over time. In this paper, we apply these modifications to deal with multi-class scenarios containing unnameable instances.

The reported results in [17] clearly demonstrate the benefit of EMOC in realistic scenarios with unnameable instances. However, the choice of Gaussian process models limits its applicability to scenarios with only several thousands of examples. In the following, we are interested in transferring their approach to linear models which will allow us to tackle significantly larger data collections.

4 EMOC for Linear Models and Large Datasets

In the following, we provide two complementary contributions to overcome the previous drawbacks of EMOC on large-scale scenarios. First of all, we present how EMOC can be applied to linear least square regression, including efficient evaluations of the criterion and efficient update rules (Section 4.1). In addition, we show how to approximate involved expectation operations to reduce computational burdens (Section 4.2).

4.1 EMOC for Linear Least Square Regression

Given the expressive power of deep learnable representations, replacing kernel methods with linear pendants can be well justifiable. For binary scenarios, a general linear

model can be written as $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ where the bias can be included into $\mathbf{w}$ by augmenting $\mathbf{x}$ with a constant dimension. The optimal solution for $\mathbf{w} \in \mathbb{R}^D$ depends on the chosen loss function.Well-known examples are logistic regression, linear SVMs, and least-square regression, which follow from minimizing the logistic loss, the hinge loss, or the quadratic loss on the training data set.Classification decisions are obtained by thresholding $f(\mathbf{x})$, *e.g.,* against zero.

For multi-class scenarios, a simple and common extension is the combination of class-specific one-vs-all models. Thereby, a binary classifier f_c is trained for each class c and the classifier with largest response on new data $\mathbf{x}$ determines the classification result. In the following, we focus on regularized least-square regression due to the resulting closed-form solutions (denoted as LSR in the remainder of the paper). Note that this directly corresponds to the linear version of Gaussian process regression as used in [17]. In this case, the hyperplanes $\mathbf{w}_c \in \mathbb{R}^D$ can be obtained as $\mathbf{w}_c = \mathbf{C}_{\text{reg}}^{-1}\mathbf{X}\mathbf{y}_c$, where the matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ holds the N training examples with feature dimension D and $\mathbf{y}_c$ is the vector of binary one-vs-all labels for class c. Furthermore, the regularized covariance matrix of the data $\mathbf{C}_{\text{reg}}$ is obtained by $\mathbf{C}_{\text{reg}} = \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_n^2\mathbf{I}$. The parameter σ_n^2 controls the degree of regularization and is related to the idea of weight decay [3].

When transferring the general EMOC criterion in Eq. (2) to LSR models, we obtain the following estimate. A detailed derivation is given in the supplementary material (see Section S1) and is purely based on applying linear algebra [22].

$$\begin{aligned}\Delta f_{mc}(\mathbf{x}') = & \frac{1}{1 + \mathbf{x}'^{\mathrm{T}}\mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'} \cdot \sum_{y' \in \mathcal{Y}} \left(p(y'|\mathbf{x}') \frac{1}{|C|} \sum_{c \in C} \left| \mathbf{w}_c^{\mathrm{T}}\mathbf{x}' - y_c' \right| \right) \\ & \cdot \frac{1}{|\mathfrak{L} \cup \mathfrak{U}|} \sum_{\mathbf{x}_j \in \mathfrak{L} \cup \mathfrak{U}} \left| \mathbf{x}_j^{\mathrm{T}}\mathbf{C}_{\text{reg}}^{-1}\mathbf{x}' \right| \quad . \end{aligned} \tag{3}$$

Since $\mathbf{C}_{\text{reg}}$ and $\mathbf{w}$ are of fixed size, the memory demand as well as required computation times remain constant for increasing training set sizes. Hence, we can evaluate the EMOC score efficiently over time (denoted by LSR-EMOC).

The second important issue for a successful active learning system is the possibility for online learning. Thereby, labeled data can be incrementally added and learning from scratch is avoided. Intuitively, this aspect gains importance for increasing dataset sizes. For the choice of one-vs-all LSR models, we can derive the following closed-form update rules which lead to efficient online learning abilities using the Sherman-Morrison-formula [23]. For a detailed derivation please also refer to Section S1:

$$\mathbf{w}_c' = \mathbf{w}_c + \mathbf{C}_{\text{reg}}^{-1}\mathbf{x}' \left(\frac{y_c' - \mathbf{x}'^{\mathrm{T}}\mathbf{w}_c}{1 + \mathbf{x}'^{\mathrm{T}}\mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'} \right) \quad , \tag{4}$$

$$\mathbf{C}_{\text{reg}}^{-1\prime} = \left(\mathbf{C}_{\text{reg}} + \mathbf{x}'\mathbf{x}'^{\mathrm{T}} \right)^{-1} = \mathbf{C}_{\text{reg}}^{-1} - \frac{\mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'\mathbf{x}'^{\mathrm{T}}\mathbf{C}_{\text{reg}}^{-1}}{1 + \mathbf{x}'^{\mathrm{T}}\mathbf{C}_{\text{reg}}^{-1}\mathbf{x}'} \quad . \tag{5}$$

We denoted with $\mathbf{w}_c'$ the new weight vector for class c after adding $\mathbf{x}'$ with corresponding binary label y_c'. Similarly, $\mathbf{C}_{\text{reg}}{}'$ denotes the updated covariance matrix. Note that the inverse of $\mathbf{C}_{\text{reg}}$ has to be computed only once and can be updated incrementally (see Eq. (5)). Since all required variables in Eq. (4) and Eq. (5) are available, the entire

update requires only $\mathcal{O}\left(D^2\right)$ operations. Similar rules based on the Cholesky decomposition [25] can be derived as well.

4.2 Approximating the Expectation Operation

Based on our previous derivations, we can directly apply the LSR-EMOC criterion to large unlabeled datasets. However, evaluating the criterion is still moderately costly due to the involved expectation operation with respect to all available data (second sum in Eq. (3)). To overcome this issue, we can approximate the expectation operation, *e.g.,* using Monte-Carlo-like sampling.

The simplest approximation is to use a randomly drawn subset $\mathcal{S}_\mathrm{r} \subset \mathfrak{L} \cup \mathfrak{U}$ when estimating model output changes. If examples in $\mathcal{S}_\mathrm{r}$ are drawn i.i.d. from all available data, we can expect that the resulting dataset statistics will be comparable. Hence, the expectation remains unchanged:

$$\Delta f_\mathrm{mc}\left(\mathbf{x}' \,|\, \mathfrak{L} \cup \mathfrak{U}\right) \approx \Delta f_\mathrm{mc}\left(\mathbf{x}' \,|\, \mathcal{S}_\mathrm{r}\right) \quad , \tag{6}$$

where we used the notation of $\Delta f_\mathrm{mc}\left(\mathbf{x}' \,|\, \mathcal{S}_\mathrm{r}\right)$ to denote that the LSR-EMOC score for example $\mathbf{x}'$ is computed using only the subset $\mathcal{S}_\mathrm{r}$. In the following, this approximation is referred to as LSR-EMOC$^{\mathrm{r}\text{-}|\mathcal{S}_\mathrm{r}|}$.

Although the property in Eq. (6) seems beneficial, a mere random selection can likely sample redundant data. We can explicitly avoid this effect by clustering all examples in advance and approximating the expectation using the set $\mathcal{S}_\mathrm{c}$ of cluster centroids only. Thereby, diversity is explicitly enforced which can be beneficial for focusing on underrepresented regions of space. We call this approximation LSR-EMOC$^{\mathrm{c}\text{-}|\mathcal{S}_\mathrm{c}|}$. Nonetheless, the equality of expectations as in Eq. (6) is no longer given.

As third alternative, we could aim at selecting an optimal subset $\mathcal{S}_\mathrm{opt} \subset \mathfrak{L} \cup \mathfrak{U}$ which leads to the closest approximation of scores:

$$\mathcal{S}_\mathrm{opt} = \underset{\mathcal{S}_* \subset \mathfrak{L} \cup \mathfrak{U}}{\operatorname{argmax}} \left| \Delta f_\mathrm{mc}\left(\mathbf{x}' \,|\, \mathfrak{L} \cup \mathfrak{U}\right) - \Delta f_\mathrm{mc}\left(\mathbf{x}' \,|\, \mathcal{S}_*\right) \right| \quad . \tag{7}$$

Unfortunately, determining the optimal subset leads to a combinatorial problem similarly complex as active learning itself. Hence, it is only theoretically feasible. In the following experimental evaluations, we provide evidence that the first two approximations are well suited for large-scale active learning tasks.

5 Experiments

In the following, we present a detailed evaluation of our LSR-EMOC approach, where we especially focus on active class discovery. Furthermore, we show the benefits of active learning for the real-world application of camera trap analysis.

5.1 Large-Scale Active Learning for Object Classification

The first part of our evaluations is concerned with the applicability of our introduced approaches to the challenging dataset MS-COCO and relevant subsets thereof. We use

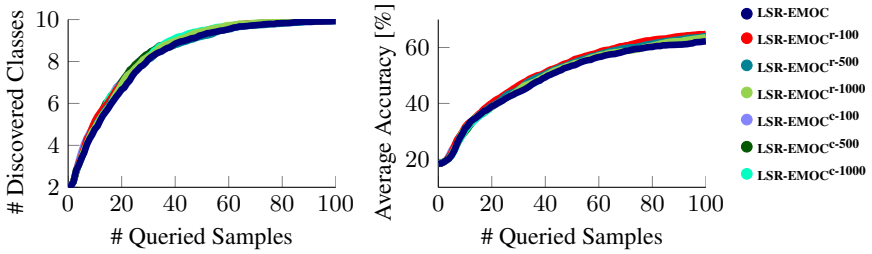


Fig. 2: Comparing approximations of LSR-EMOC on MS-COCO Animals.

the MS-COCO-full-v0.9 dataset [21] and follow the evaluation protocol of [17]. Thus, we add "noise" examples which reflect the scenario that certain examples are unnameable even for experts. In contrast to [17], we use L_2-normalized `relu7` features of the BVLC AlexNet [20]. This consistently increases the accuracy in our experiments for all active learning techniques. We optimize hyperparameters (*e.g.,* the regularization weight for LSR) on a hold-out set and keep the values fixed for all experiments.

Our approaches are compared against the baseline presented in [17], *i.e.,* EMOC on GP-regression with a linear kernel. Furthermore, we compare against several established methods: GP-Var and GP-Unc by Kapoor et al. [18], 1-vs-2 by Joshi et al. [16], and PKNN by Jain and Kapoor [15]. We use the parameter configuration schemes as in [17]. Finally, we include an upper bound which results from having all unlabeled data as labeled training examples available.

MS-COCO Animals dataset For a sanity-check evaluation, we follow [17] and use the subset of MS-COCO which corresponds to animal categories (10 categories in total). As in [17], we apply Geodesic Object Proposals [19] to obtain 3,824 samples as well as 4,574 image patches as training data and "noise" samples. Additionally, a hold-out set of 750 samples is used as validation set. Each experiment starts with 10 randomly chosen examples for each of 2 randomly chosen classes. All remaining data of the training set is used as unlabeled pool. We conduct 100 experiments with different random initializations to obtain reliable results. In each experiment, we conduct 100 queries. Learned models are evaluated after every query on a hold-out test set which is randomly drawn from the validation set and which consists of 30 samples per class.

First of all, we were interested in a comparison between EMOC of LSR models and our proposed approximations thereof. Note that LSR-EMOC leads to the same results as EMOC with GP-regression and linear kernels as used in the corresponding evaluation in [17]. Since both approaches only differ in required resource for computing scores an additional comparison regarding accuracy is not required here. Instead, we present results for LSR-EMOC and our approximation techniques in Fig. 2.

It can clearly be seen that the EMOC approach for LSR models as well as the proposed approximation techniques lead to almost identical results. Hence, we conclude that approximating EMOC calculations is possible without a notable loss in accuracy. Although the dataset is moderately small, we already obtain a speedup of 2.1 (LSR-EMOC $\approx$ 12.1s and LSR-EMOC$^{\text{r-100}}$ $\approx$ 5.7s for a whole query selection). Furthermore, it can be seen that pre-clustering of data yields no advantage over mere ran-

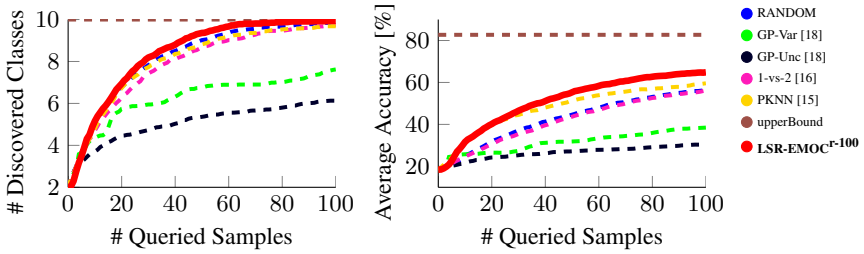


Fig. 3: Comparison of active learning methods on MS-COCO Animals.

dom selection. Hence, we conclude that the random selection should be preferred due to smaller computational costs. For the datasets used in the following evaluations, we performed similar comparisons which lead to comparable findings. Due to the lack of space, the results are shown in the supplementary material (see Section S3). In the following evaluations, we only show the fastest of our approximations, *i.e.,* random sampling with 100 samples (LSR-EMOC$^{\text{r-100}}$).

As a second experiment on the dataset, we compared our proposed method with several active learning baselines. Results are shown in Fig. 3. As can be seen, LSR-EMOC approach performs best with respect to both performance measures. Note that these results are comparable to the ones reported in [17]. However, our choice of LSR-EMOC offers scalability to larger unlabeled pools.

MS-COCO dataset Since we are interested in large-scale scenarios, we also use the entire MS-COCO dataset which can not be processed by the method of [17] in reasonable time. Image patches are obtained by using the ground truth annotations provided by the dataset. We use each box which is at least 256×256 pixel of size. Thereby, we obtain a training set of 36,212 image patches with 80 categories that consist of three to 10,632 examples. Similarly, we obtain a validation set with 46,485 patches and three to 15,986 examples per category. To keep the same ratio of unnameable instances in the unlabeled pool, we add 20,000 randomly selected patches with a maximal intersection over union score of 0.25 to any ground truth bounding box. We start with three initially known classes and 10 randomly chosen examples per category. All remaining data of the training set is used as unlabeled pool. We evaluate performances on 30 randomly selected validation samples per class. For robustness of our evaluations, we average results of nine random initialization. Since not all classes provide enough samples for the hold-out test set, we evaluate models on a fixed set which consists of 2,304 examples. In each experiment, we perform 500 query steps. All other setup parameters are kept unchanged compared to the previous section. Note further that EMOC on GP regression (*i.e.,* kernelized regularization LSR) is not longer applicable in this setting due to memory consumption and computation time. Results are shown in Fig. 4.

Again, it can clearly be seen that our method performs best with respect to the number of discovered classes. Considering accuracy, PKNN and our approach achieve comparable results. The performance drop in the beginning can be attributed to the imbalanced nature of MS-COCO. A corresponding visualization can be found in the supplementary material (see Section S2) as well as a runtime comparison (see Section S4). Note that a fast increase of accuracy can also be achieved by explicitly searching for

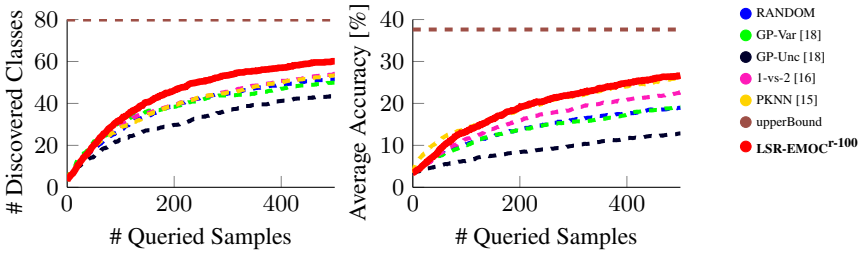


Fig. 4: Comparison of active learning methods on MS-COCO.

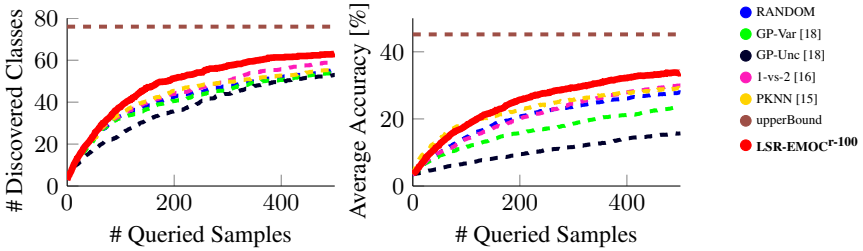


Fig. 5: Comparison of active learning methods on MS-COCO with a balanced class distribution.

rare classes as presented in [12]. However, rare class discovery is especially challenging in the presence of unnameable instances and not within the scope of this paper.

MS-COCO with a balanced class distribution We were finally interested in a comparison for a setting where classes are better balanced. Therefore, we use the previous setup and ignore categories having more than 1,000 samples in the training set. These classes are person (10,632 samples), dining table (3,979 samples), bed (1,400 samples) and cat (1,020 samples). Thereby, we obtain a training set with 19,181 samples from 76 categories and a test set with 25,060 examples. After randomly selecting 30 validation samples per class, we evaluate the methods on a test set of 2,184 samples. To augment the data with unnameable instances, 10,000 bounding boxes are randomly drawn as described previously. The remaining setup is unchanged. Results are shown in Fig. 5.

As can be seen, our method leads to superior results compared to all competitors. In direct comparison with the runner-up (PKNN), we obtain the same accuracy with only two thirds of the requested annotations. We thus conclude that our method is well suited for active learning on large-scale datasets.

5.2 Active Learning for Camera Trap Image Analysis

In the second part of our evaluations, we are interested in applying active learning to a task which arises in biodiversity assessments. Biodiversity researchers are interested in quantitative analysis of animal abundance, species composition and site occupancy. One way to treat this challenging task is to place camera traps in the wild and to record short sequences of images if any movement is detected. Thereby, researchers are faced

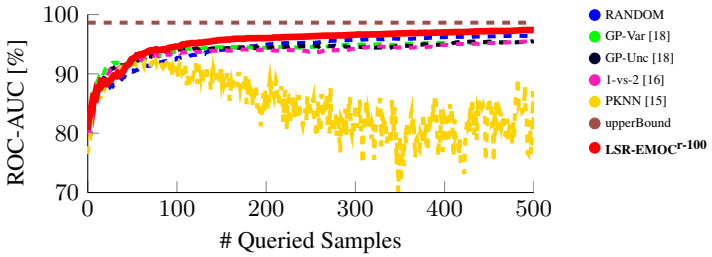


Fig. 6: Results on labeling camera trap images as either "background" or "contains-objects" with only a small number of annotations.

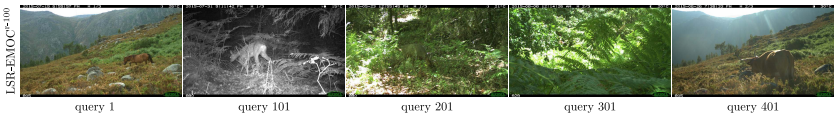


Fig. 7: Example queries from the biodiversity dataset.

with huge amounts of unlabeled data which can to date only be analyzed manually. One task is to differentiate among "background" and "contains-objects". Using active learning, we aim at training classifiers with few labeled data to solve this task.

To evaluate the benefits of active learning, we obtained a medium-scale dataset which was labeled by application experts. It consists of 2,931 frames from which 2,088 show animals. Similar to the previous evaluation, we randomly select three examples from both categories and allow 500 queries. As feature representation, we use normalized `relu7` features from the BVLC AlexNet calculated on the entire image. We repeat the random initialization 10 times to obtain reliable results. Accuracy is measured on the whole dataset after every step. Note that this is exactly the scenario which is desired by application experts: to obtain labels for the entire dataset as reliable as possible while manually labeling only few examples thereof. Results can be found in Fig. 6.

Again, our method leads to superior results compared with all competitors. In particular, we obtain the same accuracy as random selection with only 52% of required annotations. Surprisingly, the performance of PKNN is dropping over time, which we attribute to the fact that the metric learning could not cope with the variations in the data. Qualitative results can be seen in Fig. 7 as well as in Section S5. According to application experts, the results already lead to a valuable reduction of annotation costs.

6 Conclusion

We presented an active learning technique able to cope with large-scale data. Our technique is based on the principle of expected model output changes and builds on two complementary aspects: the effectiveness of linear models as well as careful approximations of necessary computations. We provided empirical evidence that our approach is capable of performing well even on challenging imbalanced datasets. Furthermore, we presented real world experiments for biodiversity data analysis which finally show the applicability and effectiveness of our method.

References

1. Alajlan, N., Pasolli, E., Melgani, F., Franzoso, A.: Large-scale image classification using active learning. IEEE Geoscience and Remote Sensing Letters 11(1), 259–263 (2014)
2. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. Journal of Machine Learning Research (JMLR) 5, 255–291 (Dec 2004)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics, Springer (2006)
4. Bouneffouf, D.: Exponentiated gradient exploration for active learning. Computers 5(1), 1 (2016)
5. Cai, W., Zhang, Y., Zhou, S., Wang, W., Ding, C.H.Q., Gu, X.: Active learning for support vector machines with maximum model change. In: European Conference Machine Learning and Knowledge Discovery in Databases (ECML, PKDD). pp. 211–226 (2014)
6. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning 15(2), 201–221 (1994)
7. Demir, B., Bruzzone, L.: A novel active learning method in relevance feedback for content-based remote sensing image retrieval. IEEE Transactions on Geoscience and Remote Sensing 53(5), 2323–2334 (2015)
8. Ertekin, S., Huang, J., Bottou, L., Giles, L.: Learning on the border: active learning in imbalanced data classification. In: ACM Conference on information and knowledge management. pp. 127–136 (2007)
9. Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Labeling examples that matter: Relevance-based active learning with gaussian processes. In: German Conference on Pattern Recognition (GCPR) (2013)
10. Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: Active learning with expected model output changes. In: European Conference on Computer Vision (ECCV). pp. 562–577 (2014)
11. Fu, C., Yang, Y.: A batch-mode active learning SVM method based on semi-supervised clustering. Intelligent Data Analysis 19(2), 345–358 (2015)
12. Haines, T.S.F., Xiang, T.: Active rare class discovery and classification using dirichlet processes. International Journal of Computer Vision (IJCV) 106(3), 315–331 (2014)
13. Hoi, S.C., Jin, R., Lyu, M.R.: Large-scale text categorization by batch mode active learning. In: ACM International Conference on World Wide Web. pp. 633–642 (2006)
14. Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. In: Neural Information Processing Systems (NIPS). pp. 892–900 (2010)
15. Jain, P., Kapoor, A.: Active learning for large multi-class problems. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 762 –769 (2009)
16. Joshi, A., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2372 –2379 (2009)
17. Käding, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Active learning and discovery of object categories in the presence of unnameable instances. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4343–4352 (2015)
18. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. International Journal of Computer Vision (IJCV) 88, 169–188 (2010)
19. Krähenbühl, P., Koltun, V.: Geodesic object proposals. In: European Conference on Computer Vision (ECCV), pp. 725–739 (2014)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS). pp. 1097–1105 (2012)

21. Lin, T., Maire, M., Belongie, S., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
22. Plackett, R.L.: Some theorems in least squares. Biometrika 37(1/2), pp. 149–157 (1950)
23. Press, W.H.: Numerical recipes 3rd edition: The art of scientific computing. Cambridge university press (2007)
24. 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)
25. Seeger, M.: Low rank updates for the cholesky decomposition. Tech. rep., University of California, Berkeley (2004)
26. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison (2009)
27. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research (JMLR) 2, 45–66 (2002)
28. Vezhnevets, A., Buhmann, J.M., Ferrari, V.: Active learning for semantic segmentation with expected change. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
29. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. International Journal of Computer Vision (IJCV) 113(2), 113–127 (2014)