A. Freytag, E. Rodner, and J. Denzler
Selecting Influential Examples: Active Learning with Expected Model Output Changes. ECCV 2014


# -Supplementary Material- 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.** The following document gives additional information with respect to the paper *Selecting influential examples: Active Learning with Expected Model Output Changes*. Details for five aspects are presented: (i) reasons and experimental validation for choosing label regression (Sect. S1), (ii) experimental setup for visualizing active learning, which led to Fig. 2 in the main submission (Sect. S2), (iii) an in-depth comparison of active learning techniques with respect to desired properties (Sect. S3), (iv) detailed information regarding the experimental setup used in the main submission (Sect. S4), and (v) a visualization of queried images for several one-vs-all tasks on Caltech256 (Sect. S5). The provided information is not necessary to understand the main paper.

## S1 Why Did You Choose Label Regression?

Classification with GP models is often done with approximate inference methods like Laplace approximation or expectation propagation [6], which allow for noise models theoretically more suitable for classification. In our case, we used label regression [5], which applies GP regression directly to discrete classification labels. This technique is very much related to least-squares SVM and a common classification strategy [5,4]. Furthermore, we also evaluated the performance of label regression compared to Laplace approximation and expectation propagation. In particular, we learned classification models for all 100 binary ImageNet tasks also used in the previous experiments of the main paper (30 training examples used for each category) with all three methods. The approximate inference methods (using the classification model in Eq. (8)) performed significantly worse with **83.29%** AUC (Laplace) and **83.09%** AUC (EP) than label regression with **85.55%** AUC ($t$-test, $p < 10^{-9}$). This performance gap is even more severe with less training examples. However, it should be emphasized again that our EMOC strategy introduced in the main paper can be applied in general also to these approximate inference methods but without the efficient model updates presented in Sect. 4 of the submission.

## S2 Experimental Setup for Visually Inspecting Active Learning

As described in the main paper, we have been interested in understanding advantages and disadvantages of state-of-the-art active learning techniques on a fundamental level. Therefore, we designed a controlled setup offering several burdens each technique is focused with:

- **Small labeling budget**: compared to the overall number of unlabeled samples (120 samples to choose from), we restricted the labeling budget to only 9 points to get an impression of what the different techniques focus on most.
- **Existence of outliers**: we added 9 outliers clearly far off the main sample distribution to evaluate the resistance to outliers of every technique.
- **Initially unknown clusters**: in order to inspect the ability of discovering relevant new clusters in feature space, we added 2 initially unknown clusters to the unlabeled pool.
- **Sub-optimal initial decision boundaries**: since the initial decision boundary is far from being optimal, we can visually check whether a technique aims for improving especially the current boundary to make decisions in known regions more reliable.

In order to easily visualize the process of active learning, we restricted the feature space to two dimensions represented by $x$ and $y$ coordinates. An RBF-kernel serves as ad-hoc choice for measuring sample similarity. In the following visualizations, data already labeled is indicated with white diamonds and white crosses for samples of positive and negative class, respectively. Unlabeled data is plotted in white dots and thickness of dots corresponds to their score obtained by the active learning criterion currently inspected (thickest dots are preferred). For every iteration and method, the unlabeled sample to be queried next is colored in magenta. Current classification scores for the entire input space are color coded, with red colors and blue colors indicating tendency to the positive and negative class, respectively. Green colors correspond to regions where estimations for both classes are currently on par. Together with the source code developed, we will make the data and evaluation protocol publically available upon acceptance.

## S3 Summary of Findings

We applied several state-of-the-art strategies for active learning from different general techniques to the 2D problem introduced before (see Sect. (2) in the main submission for an overview of related work on active learning and details of the strategies used here). A brief summary of our findings is depicted in Table 1.

In the following, we outline detailed observations we made regarding every analyzed technique. Results are visualized in Fig. 2 and close-up inspections of the actual region of interest in feature space are shown in Fig. 3, respectively. Note that we excluded visual results for passive learning (random queries) due to its non-deterministic behavior.

**Table 1.** Results of visually analyzing active learning techniques with respect to desired properties.

| Strategy | Resistant to outliers | Discovers new clusters | Improves decision boundary | Confirmable on real-world data |
|---|---|---|---|---|
| Random | | | | ✓ |
| Predictive variance [4] | | (✓) | | ✓ |
| Classification uncertainty [4] | | (✓) | | ✓ |
| Model change [2] | ✓ | | ✓ | ✓ |
| Reduction of classif. error [7] | ✓ | ✓ | ✓ | |
| **EMOC strategy (Ours)** | ✓ | ✓ | ✓ | ✓ |

**Predictive variance [4] – misled by outliers** By design, querying samples being maximally distant from all previously known data is a purely explorative strategy. This clearly reflects in the samples chosen (first row in Fig. 2), which are by no means related to the actual problem of interest. We therefore conclude that this strategy is useful in scenarios, where the available data is known to contain no outliers, and a rapid coverage of all possible inputs is needed. However, for at least slightly disturbed data collections, focusing on yet unexplored regions in feature space might waste expensive labeling budget as can be easily seen in the points picked, and consequently can perform even worse than passive learning (random selection, see also Sect. 6.1 in the main submission)

**Classification uncertainty [4] – surprisingly misled by outliers too** Although looking for samples with highest classification uncertainty intuitively seems to look for decision boundaries, the picked samples for the 2D experiment are in fact the same as chosen by (GP-var). While being counter-intuitive on first sight, the current model naturally has no accessible information for outliers, and it consequently focuses on first treating those regions, too. In other words, unexplored regions in space can be seen as 'a huge decision boundary' due to the zero mean prior. Note that this effect can also happen for different classification models, *e.g.*, SVMs as used by [8].

**Model change [2] – a focus possibly too strong on current decision boundaries** For the previous two strategies, we did not had to investigate the close-up visualizations at all, since queried samples had been far away from initial training data, thereby only visible in Fig. 2. In contrast, the model change criterion results in samples close to initially known regions, as can be seen in Fig. 3, third row. Here, we clearly observe two properties: (i) the strategy seems to be incapable of finding new clusters of data, and (ii) it prefers samples surprisingly similar to labeled ones (*e.g.*, see Iter. 2 and Iter. 3). Frankly, this observation took us by surprise, since [2] introduced the method as "implicit balancing between exploration and exploitation", and their theoretical derivation was confirmed by experiments both on artificial and on real-world data. However, at least for the 2D setup, we noticed the balancing to be heavily biased towards exploitation.

**Reduction of classif. error [7] – a fair balance of exploration and exploitation** From Fig. 2, it is evident that empirical risk minimization seems not to be tricked by outliers at all. Furthermore, we can see a well-balanced behavior of exploring new clusters (*e.g.*, Iter. 1, Iter. 2, and Iter. 3 in Fig. 3, row 4) and improving current decision boundaries (*e.g.*, Iter. 4 and Iter. 5). For the designed 2D example, this strategy seems to perfectly offer all abilities one might desire for an active learning criterion. This is the more remarkable, since we could not confirm this observation on real-world data (see Sect. 6.1 in the main submission). We believe this origins from real-world class distributions being far more complex, leading to bad initial label estimates, followed by bad empirical risk estimates. In direct consequence, we argue that heading for "decisions turning for the better" is often too ambitious, since "the better" can be wrongly estimated too easily, especially with only few labeled data available.

**EMOC strategy (Ours) – the best of both worlds and confirmable on real-world data** Similar to the results of empirical risk minimization, looking for expected model output changes turns out to be resistant to outliers (row 5 in Fig. 2), and perfectly discovers new relevant clusters in data while simultaneously improves current decision boundaries (row 5 in Fig. 3). Furthermore, our technique circumvents real-world problems of [7] by looking for "decisions turning" – without being too picky about the "for the better" aspect (see Sect. 6.1 in the main submission).

## S4 Experimental Setup for Main Submission

**Datasets** Experimental results presented in the main submission (see Sect. 6) are obtained on the two established dataset Caltech-256 [3] and ImageNet [1].

**Sampling of Classes and Samples** For the majority of experiments, 100 sets of 10 classes have been drawn uniformly from the available classes. With every set, 10 random partitions were created for reliable results. Thus, for ImageNet, the official train set was used to uniformly sample 1 training image per class, and 99 remaining samples per class where used as unlabeled examples. The separate test set was used to obtain 50 samples per class for accuracy evaluation. On Caltech-256, we split data per class into 1 sample for training, 30 hold-out samples for testing, and the remaining samples Thus, 1 sample per class was picked for training, 30 additional samples served as hold-out set for testing, and all remaining samples were collected in an unlabeled pool.

**Features** Image representations were based on publicly available bag-of-visual-words features of ILSVRC 2010, resulting in 1 000 dimensional L1-normalized histograms.

**Kernel Function** A histogram intersection kernel served as similarity measure, since its superior results in previous works over standard kernels on histograms, *e.g.*, RBF-kernels. In addition, no hyperparameter optimization needs to be done.

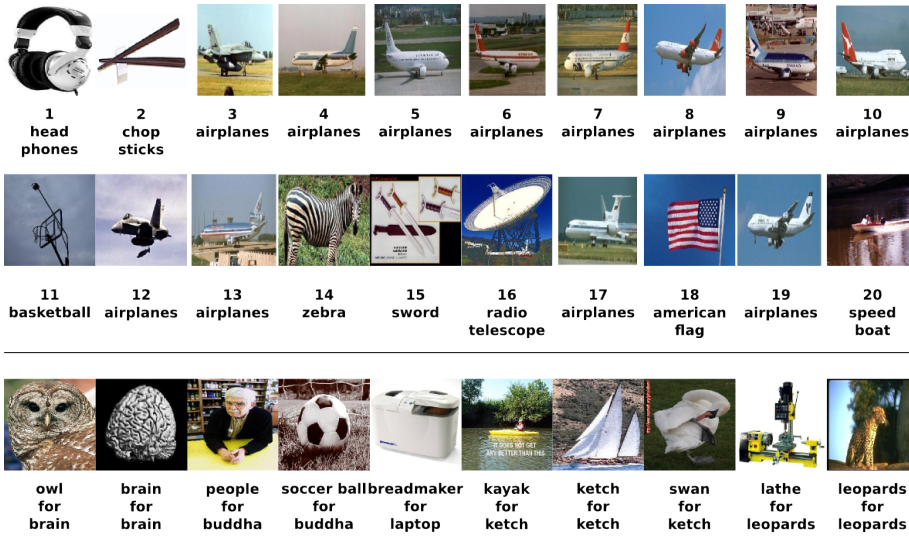


**Fig. 1.** Images picked in 1-vs-all experiments on Caltech-256. *Top*: given are the first 20 queries made by our FastEMOC technique for airplane as positive class. *Bottom*: Some interesting queries made by our FastEMOC technique for several classes.

**Model Regularization** For model regularization, the noise level was set to $\sigma_n^2 = 0.1$ in all our experiments. We also tested with initial optimization or optimization after 5 or 10 queries, but found no superior behavior with respect to recognition rates.

**Accuracy Determination** Since we are interested in active learning of class detectors, all scenarios are binary ones. Thus, we evaluated accuracy after every query using the area under ROC curves, to be independent of threshold determination.

**One-vs-all-Experiments (Fig. 4 in main submission)** We additionally conducted 1-vs-all experiments as suggested during the process of reviewing, following the evaluation setup of [4]. Therefore, all classes of Caltech-256, which have already been present in Caltech-101, served ones as positive class, and all remaining 255 classes as negatives. The picked class IDs are [15, 20, 22, 36, 52, 57, 63, 64, 66, 91, 100, 102, 114, 121, 123, 127, 129 134, 140, 145, 172, 179, 201, 204, 230, 235, 240, 251, 252, 253]. For every of the 30 resulting 1-vs-all settings, we followed the exact setup of [4], and we started with 1 example per class for training ( 1 positive, 255 negatives). Additional 10 samples per class served as held-out test set. Every binary task for randomly initialized for 10 and results are averaged. Gain in accuracy over passive learning (random) is plotted after a total of 20 queries.

## S5 Visualizing Queried Images of One-vs-all Tasks

Finally, we visualize images actively selected by our FastEMOC technique for some of the aforementioned one-vs-all scenarios. More precisely, we plotted the first 20 queries for a one-vs-all task on Caltech256's airplane category in the top part of Fig. 1. As can be seen, our strategy nicely balances querying positive and negative samples, and especially refining class boundaries with mixed-up categories (*e.g.*, zebra, speed boat, and american flag). In the bottom part of Fig. 1, we additionally plotted query results which we found interesting during writing the paper, *e.g.*, an owl, which was queried during a task with brains as positive class, or a soccer ball, which likely was queried to better differentiate between compact buddhas and visually similar round objects.

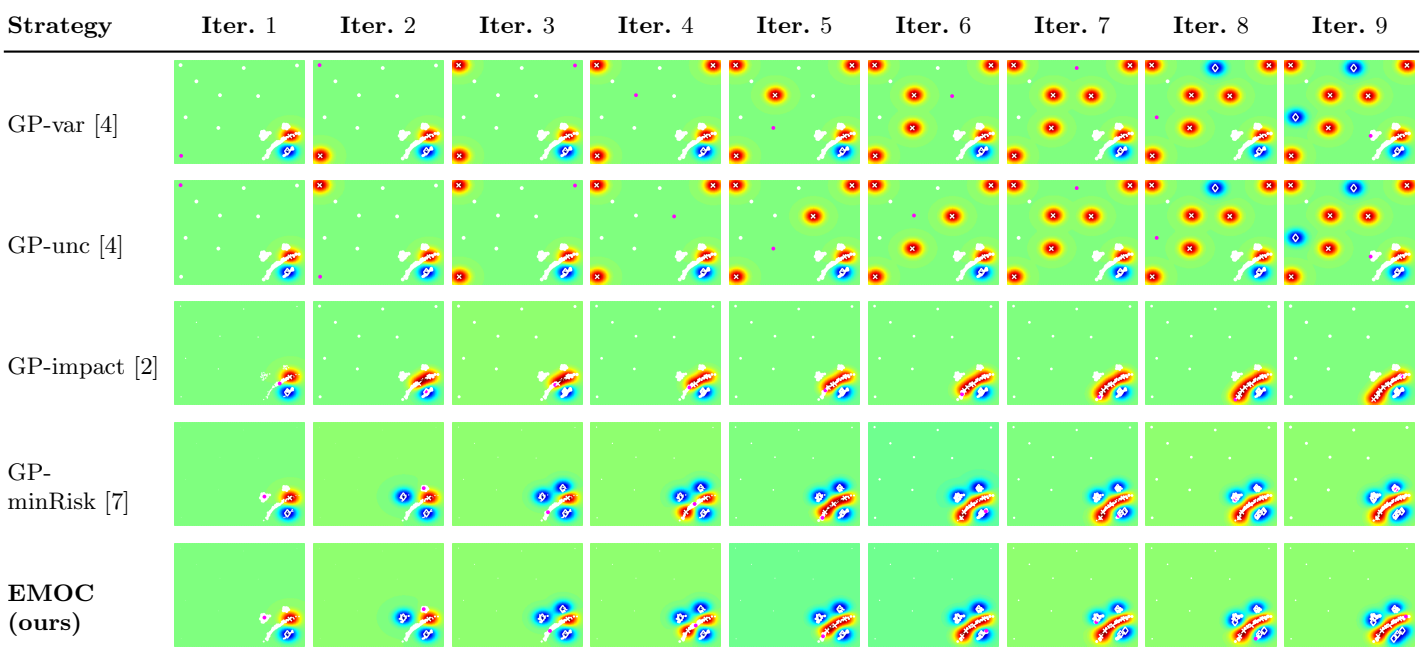


**Fig. 2.** Inspecting active learning on a 2D toy example. The figure is best viewed in color and by zooming in. See text above for further explanations.

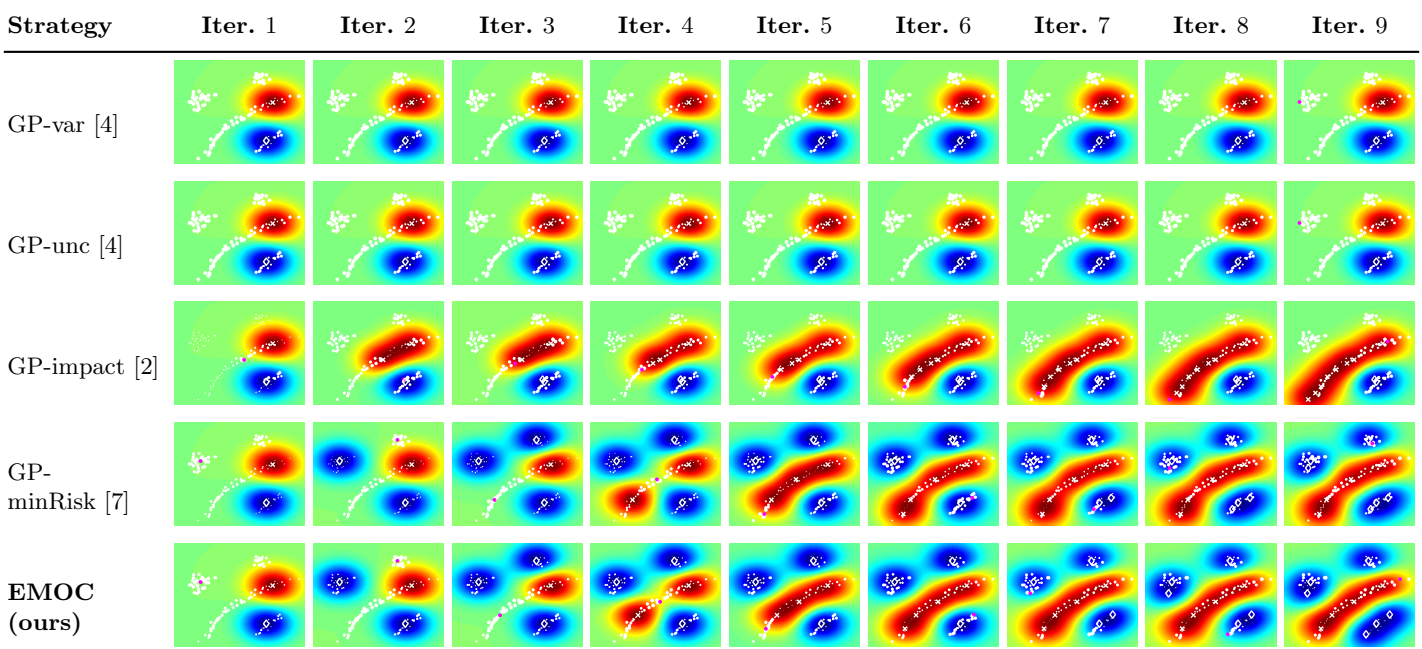


**Fig. 3.** Close-up inspections of active learning visualizations presented in Fig. 2. Shown here is the bottom right corner, where the data being problem relevant is located. The figure is best viewed in color and by zooming in. See text above for further explanations.

## References

1. 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)
2. 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). pp. 282–291 (2013)
3. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007)
4. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. International Journal of Computer Vision (IJCV) 88, 169–188 (2010)
5. Nickisch, H., Rasmussen, C.E.: Approximations for binary gaussian process classification. Journal of Machine Learning Research 9(10) (2008)
6. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning, The MIT Press, Cambridge, MA, USA (01 2006)
7. 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)
8. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research (JMLR) 2, 45–66 (2002)