Labeling Examples That Matter: Relevance-Based Active Learning with Gaussian Processes

Alexander Freytag[1], Erik Rodner[1,2], Paul Bodesheim[1], and Joachim Denzler[1]

[1] Computer Vision Group, Friedrich Schiller University Jena, Germany
http://www.inf-cv.uni-jena.de
[2] UC Berkeley ICSI & EECS, United States

Abstract. Active learning is an essential tool to reduce manual annotation costs in the presence of large amounts of unsupervised data. In this paper, we introduce new active learning methods based on measuring the impact of a new example on the current model. This is done by deriving model changes of Gaussian process models in closed form. Furthermore, we study typical pitfalls in active learning and show that our methods automatically balance between the exploitation and the exploration trade-off. Experiments are performed with established benchmark datasets for visual object recognition and show that our new active learning techniques are able to outperform state-of-the-art methods.

1 Introduction

The amount of visual data available on the Internet is tremendous and still rapidly increasing, for an example, around three million images are uploaded to Facebook every day. When annotated properly, such image collections are powerful sources for building and improving visual recognition algorithms that learn object and category models. The ImageNet project [6] for example used Internet search engines and manual filtering done by Amazon Mechanical Turk workers to provide labeled images for more than $10,000$ different classes and semantic concepts. However, the majority of available images on the net is biased towards very common and similar object instances (*e.g.*, standard white mugs in different resolutions). If we automatically select relevant sample images and manually ensure their correct labeling, we would generate annotated visual data more efficiently. Furthermore, for real-world robotics scenarios, limiting ourselves to a fixed number of classes beforehand is likely to fail due to the large variety of application-specific semantic concepts.

Due to these reasons, we consider active learning in this paper, which allows for actively obtaining labels from human annotators for samples that are likely to be important to improve the current classifier. In particular, given a set of unlabeled examples, an active learning algorithm has to pick a specific example, which is then labeled and added to the training set. The goal of active learning in this case is to save annotation budget and to allow for steeper learning curves that lead to higher recognition performances with fewer labeled training examples.

J. Weickert, M. Hein, and B. Schiele (Eds.): GCPR 2013, LNCS 8142, pp. 282–291, 2013.

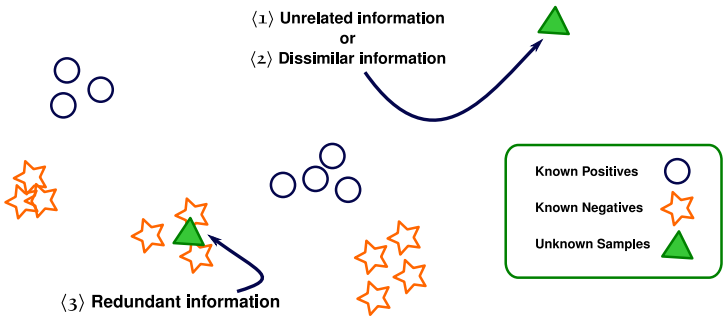

Fig. 1. Typical pitfalls in active learning scenarios: queried examples are either possible outliers (*unrelated* ⟨1⟩ or *dissimilar* ⟨2⟩ to the current distribution), or the samples will likely to be *redundant* ⟨3⟩. Our methods implicitly balance both extrema automatically by selecting samples that are useful (Given numbers reflect terms in Theorem 1).

The difficulty of active learning is the critical trade-off between exploration and exploitation, *i.e.*, is it beneficial to explore the feature space by selecting examples with a large distance to the current training set or should we select examples nearby with the risk that they are likely to be redundant (see Fig. 1)? Our approach directly tackles this trade-off by selecting examples that do have an impact on the classification model. For classification, we use Gaussian process (GP) models [13], which provide probabilistic non-linear classifiers and achieve comparable performance to SVM models. One of the advantages of the GP regression framework is that we can update the model in closed form for the case a new training example is added, which is derived in this paper. The update formulas allow us to measure the impact of an example on the model in various ways leading to new active learning methods. We also show the underlying relation to other approaches, like the one in [10], and compare our results to them in several experiments. The contributions of this paper are as follows:

1. **Model change for GP**: We derive update formulas for GP regression which give interesting insights into what is going on when new examples are added to the current model.
2. **Active learning strategies based on expected change**: We derive two strategies for binary scenarios based on the impact of a new example on the current classification model. Both strategies are shown to be suitable for active learning in several scenarios.
3. **Robustness with respect to initial set size**: In contrast to state-of-the-art query strategies, our methods are able to significantly reduce the labeling effort even in the presence of extremely few initially known samples.

We briefly review related approaches and active learning with Gaussian processes in Sect. 2 and 3. Closed-form model updates for GP regression are presented in detail and our active learning methods are derived in Sect. 4. Experimental results are given in Sect. 5 highlighting the benefits of our introduced methods in several active learning scenarios. A summary and outlook concludes the paper.

2 Related Work

Definition of Active Learning. The pipeline in an active learning scenario is as follows: (1) collect a large number of unlabeled examples $\mathcal{U} = \{\widehat{\boldsymbol{x}}^{(1)}, \ldots, \widehat{\boldsymbol{x}}^{(m)}\}$, (2) hand them over to a query function $\mathcal{Q}$ to compute a score for every sample, and (3) choose a sample $\widehat{\boldsymbol{x}}^{(i)} \in \mathcal{U}$ with best score to label it and provide it as an additional training example for a classifier. State-of-the-art methods can be divided into those that directly aim at minimizing the expected future classification error of the classifier [8,14] or alternatively try to reduce the space of hypotheses consistent with the data known so far as fast as possible [4,15,16,17,10].

Combining Exploration and Exploitation. In active learning scenarios, the decision has to be made whether to spend some (annotation) budget in exploring unseen regions of the feature space or whether to alternatively refine the current decision boundaries in order to improve separability of known classes. While the first aspect is denoted as *exploration*, the second one is usually referred to as *exploitation*. However, relying on a single active learning aspect is unfeasible for many if not all real-world applications [12,7]. Depending on the dataset as well as the actual problem and the currently known samples, both aspects need to be taken into account but are not equally important. Baram et al. [1] propose selecting the current criterion from a pool of query strategies using a multi-armed bandit. Osugi et al. [12] flip a biased coin to decide between pure exploration and pure exploitation. Furthermore, they update the probability depending on the change of the current model but without considering the actual success of the query. Ebert et al. [7] propose switching between different explorative and exploitative methods using a Markov decision process formulation and update the state transitions online as well. All of the previously presented approaches aim at explicitly selecting the currently most valuable query strategy. In contrast, our approach implicitly combines exploration and exploitation without the necessity of a selection step or a decision model.

Work Most Similar to Our Approach. Cebron and Berthold [5] propose using a linear combination of an explorative and an exploitative strategy. However, the linear weighting has to be defined explicitly and in addition the exploration method is "out-of-the-box" without a clear relation to the classification model used. Kapoor et al. [10] propose combining expected mean and variance of a Gaussian process classifier by dividing both scores, which can only be justified heuristically. In contrast to both approaches, our techniques are derived from the underlying classification model without any parameters to tune. Similar in spirit is the approach of Vezhnevets et al. [18], where the expected change of a conditional random field after including a labeled example is taken as active learning score. To compute the expected change, assumptions and simplifications have to be made to handle computational costs. For Gaussian process models, Bodesheim et al. [3] introduced an approximation of the expected model change only based on the change of the model output for the new sample. In contrast to both techniques, our approach does not need any assumption or simplification to assess the change of the model.

3 Active Learning with Gaussian Processes

It has been shown that Gaussian process models are useful for active learning [16], especially in visual object categorization [10]. Using the Gaussian process regression framework, binary class labels $y_i \in \{1, -1\}$ of samples $\boldsymbol{x}^{(i)}$ are treated as continuous values which are assumed to be outputs of a latent function f. Following a Gaussian noise model, function values are assumed to be disturbed by white Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_{\mathsf{n}}^2)$ leading to $y_i = f(\boldsymbol{x}^{(i)}) + \epsilon$. Latent function values $\boldsymbol{f}$ for any finite set of samples $\mathbf{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$ are modeled to be jointly Gaussian with a zero mean function and a covariance function κ, *i.e.*, $\boldsymbol{f} \sim \mathcal{GP}(\mathbf{0}, \kappa(\mathbf{X}, \mathbf{X}))$. Given a set of training samples, marginalization over latent function values leads to a Gaussian posterior distribution for the label of a test sample $\boldsymbol{x}^*$, where predictive mean and variance can be computed in closed form:

$$\mu_*(\boldsymbol{x}^*) = \boldsymbol{k}_*^{\mathrm{T}} \left(\mathbf{K} + \sigma_{\mathsf{n}}^2 \mathbf{I}\right)^{-1} \boldsymbol{y} = \boldsymbol{k}_*^{\mathrm{T}} \boldsymbol{\alpha} \ , \tag{1}$$

$$\sigma_*^2(\boldsymbol{x}^*) = k_{**} + \sigma_{\mathsf{n}}^2 - \boldsymbol{k}_*^{\mathrm{T}} \left(\mathbf{K} + \sigma_{\mathsf{n}}^2 \mathbf{I}\right)^{-1} \boldsymbol{k}_* = \sigma_{f_*}^2 + \sigma_{\mathsf{n}}^2 \ . \tag{2}$$

Here, we use $\mathbf{K}$, $\boldsymbol{k}_*$, and k_{**} to denote the kernel matrix containing covariances of training samples, the kernel vector containing covariances between the training samples and the test sample, and the covariance of the test sample to itself, respectively. The sign of the predictive mean is typically used for classification. Note that within this paper, we focus on binary classification as done by [10], which offers a large variety of scenarios to be tackled. Object detection methods for example rely on this setup and learn from positive samples of a specific object class and lots of negative data from arbitrary other classes and background textures. For multi-class settings, simple techniques like using the score difference between the best and the second best class usually work well.

As proposed in [10], samples from a huge pool of unlabeled data can be queried using one of the following three strategies. Selecting samples by minimum absolute predictive mean: $\mathcal{Q}_{\mu_*}(\mathcal{U}) = \operatorname{argmin}_{\widehat{\boldsymbol{x}}^{(i)} \in \mathcal{U}} \ |\mu_*(\widehat{\boldsymbol{x}}^{(i)})|$ is an exploitative strategy, since such samples are near the decision boundary of the classification model. Samples that are far from already known data can be selected via a large predictive variance: $\mathcal{Q}_{\sigma_*^2}(\mathcal{U}) = \operatorname{argmax}_{\widehat{\boldsymbol{x}}^{(i)} \in \mathcal{U}} \ \sigma_*^2(\widehat{\boldsymbol{x}}^{(i)})$ to explore the feature space. As a combination of exploitation and exploration it is suggested to query samples based on uncertainty defined by: $\mathcal{Q}_{\mathrm{unc}}(\mathcal{U}) = \operatorname{argmin}_{\widehat{\boldsymbol{x}}^{(i)} \in \mathcal{U}} \ \frac{|\mu_*(\widehat{\boldsymbol{x}}^{(i)})|}{\sqrt{\sigma_*^2(\widehat{\boldsymbol{x}}^{(i)})}}$.

However, rather than defining a combination of exploration and exploitation heuristically, we present strategies that are based on theoretical foundations of Gaussian process regression models and corresponding model changes. These strategies are derived in the next section.

4 Relevance-Based Active Learning

In an active learning setup, one seeks to query examples that are most valuable with respect to the given task. As presented in Sect. 2, several approaches exist to define what is *valuable*, *e.g.*, examples that are maximally unknown or those

leading to the largest estimated gain of recognition rates. However, when we pick a new example to be labeled, we should ensure that the labeling process will not be in vain, *i.e.*, that the chosen example will be taken into account when building the new model.

Due to the representer theorem, classification scores of many popular classification models can be computed as a weighted sum of similarities between test example $\boldsymbol{x}^*$ and all training data, *i.e.*, $s(\boldsymbol{x}^*) = \boldsymbol{k}_*^{\mathrm{T}}\boldsymbol{\alpha}$. Common examples are support vector machines (SVM) or Gaussian processes (GP), which have been briefly reviewed in the previous section. If an entry α_i is almost zero, the influence of the corresponding example vanishes. If we estimate or even predict the weight $\bar{\alpha}_{n+1}$ of the updated model parameters $\bar{\boldsymbol{\alpha}}$ for a new example, we could only take examples into account for labeling, that have a large influence and are therefore worth being labeled.

4.1 Computing Impact Using Gaussian Processes

In contrast to SVMs, where $\bar{\boldsymbol{\alpha}}$ is only available after convex optimization [19], we can indeed compute $\bar{\boldsymbol{\alpha}}$ for GP in closed form. Let us have a closer look on the change of the weight vector when a new example $\boldsymbol{x}^*$ is added. The new kernel matrix $\bar{\mathbf{K}}$ after selecting $\boldsymbol{x}^*$ with binary label y_* is given by:

$$\bar{\mathbf{K}} = \begin{bmatrix} \mathbf{K} + \sigma_{\mathsf{n}}^2 \cdot \mathbf{I} & \boldsymbol{k}_* \\ \boldsymbol{k}_*^{\mathrm{T}} & \kappa(\boldsymbol{x}^*, \boldsymbol{x}^*) + \sigma_{\mathsf{n}}^2 \end{bmatrix} \quad . \tag{3}$$

Based on the following theorem[1] , we can compute new weights $\bar{\boldsymbol{\alpha}}$ in closed form.

Theorem 1 (Closed form update of GP regression weights). *Let $\boldsymbol{x}^*$, y_*, $\mathbf{K}$, $\boldsymbol{k}_*$, k_{**}, $\boldsymbol{\alpha}$, σ_{n}^2, and $\sigma_{f_*}^2$ be as previously defined. Then we can compute the weight vector $\bar{\boldsymbol{\alpha}}$ after adding $\boldsymbol{x}^*$ to the training set as follows:*

$$\bar{\boldsymbol{\alpha}} = \bar{\mathbf{K}}^{-1} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha} \\ 0 \end{bmatrix} \quad + \quad \underbrace{\frac{1}{\sigma_{f_*}^2 + \sigma_{\mathsf{n}}^2}}_{\langle 1 \rangle} \underbrace{\begin{bmatrix} \left(\mathbf{K} + \sigma_{\mathsf{n}}^2 \cdot \mathbf{I}\right)^{-1} \boldsymbol{k}_* \\ -1 \end{bmatrix}}_{\langle 2 \rangle} \underbrace{\left(\boldsymbol{k}_*^{\mathrm{T}}\boldsymbol{\alpha} - y_*\right)}_{\langle 3 \rangle} . \tag{T1}$$

The three factors in Eq. (T1) can be nicely interpreted. The first term $\langle 1 \rangle$ states that if a new example is *unrelated* to the distribution of current training data, *i.e.*, the predictive variance is large indicating a possible outlier [11], the weight $\bar{\alpha}_{n+1}$ for the example as well as the overall model change $\triangle = \left|\left|\bar{\boldsymbol{\alpha}} - (\boldsymbol{\alpha}^{\mathrm{T}}, 0)^{\mathrm{T}}\right|\right|$ will be small. Expression $\langle 2 \rangle$ can also be interpreted in a similar fashion as a weighted Parzen estimate and we observe that if a new example is again *dissimilar* with respect to currently known training examples, the overall model change $\triangle$ will again be small. Finally, in $\langle 3 \rangle$ we notice that if *redundant* information is to be added, *i.e.*, the new label can already be explained given the current model, the new weight $\bar{\alpha}_{n+1}$ as well as the overall model change $\triangle$ will be small as well. A visual explanation of this behavior is shown in Fig. 1.

[1] A complete proof based on block matrix inversion [2, p. 117] is given in the suppl. material.

4.2 Derived Active Learning Strategies

We derive two new GP query methods from the previous result. Whereas the first strategy only considers the resulting weight for a new example, the second takes the overall model change into account.

From Theorem 1 we know that the new entry of the updated alpha vector only depends on the first and third term. However, the ground-truth label y_* is not known before querying it, and we have to make assumptions based on the information currently available. In absence of further information, we choose the most pessimistic estimate of model change for a given example $\widehat{\boldsymbol{x}}$:

$$\mathcal{Q}_{\text{weight}}(\mathcal{U}) = \underset{\widehat{\boldsymbol{x}}^{(i)} \in \mathcal{U}}{\operatorname{argmax}} \min_{y^{(i)} \in \{-1,1\}} \frac{|\mu_*(\widehat{\boldsymbol{x}}^{(i)}) - y^{(i)}|}{\sigma_f^2(\widehat{\boldsymbol{x}}^{(i)}) + \sigma_{\mathsf{n}}^2} \ . \tag{4}$$

This strategy can be interpreted as an implicit balancing between exploitative methods (enumerator) and explorative methods (denominator) [5,7].

As mentioned earlier, the second strategy will also take the overall model change into account. The underlying assumption is that a sample, which would heavily affect the current model even with the most plausible label, is worth being labeled. We make use of Theorem 1 and arrive at the following:

$$\mathcal{Q}_{\text{impact}}(\mathcal{U}) = \underset{\widehat{\boldsymbol{x}}^{(i)} \in \mathcal{U}}{\operatorname{argmax}} \min_{y^{(i)} \in \{-1,1\}} \left|\left| \Delta\boldsymbol{\alpha}^{(i)} \right|\right|_1 \tag{5}$$

$$= \underset{\widehat{\boldsymbol{x}}^{(i)} \in \mathcal{U}}{\operatorname{argmax}} \min_{y^{(i)} \in \{-1,1\}} \left|\left| \frac{|\mu_*(\widehat{\boldsymbol{x}}^{(i)}) - y^{(i)}|}{\sigma_f^2(\widehat{\boldsymbol{x}}^{(i)}) + \sigma_{\mathsf{n}}^2} \cdot \begin{bmatrix} \left(\mathbf{K} + \sigma_{\mathsf{n}}^2 \cdot \mathbf{I}\right)^{-1} \boldsymbol{k}_*^{(i)} \\ -1 \end{bmatrix} \right|\right|_1 \tag{6}$$

We see in the next section as well as in the experimental results (Sect. 5) that both methods implicitly adapt the amount of exploration and exploitation leading to superior learning rates compared to state-of-the-art techniques.

4.3 Trade-Off between Exploration and Exploitation

In the following, we analyze the behavior of the derived strategies on a synthetic 1D toy example. As visualized in Fig. 2, we only know two positive and a single negative example. Examples are represented based on a single 1D feature value, and similarity is measured using a standard RBF kernel. Both plots show the scores of the different GP based query strategies for a broad range of possible inputs. Note that we do not include $\mathcal{Q}_{\text{unc}}$ into the figure for the sake of simplicity. Apart from this, its behavior is pretty similar to $\mathcal{Q}_{\mu_*}$.

For computing the left plot, the bandwidth parameter of the RBF kernel σ_{RBF} was set to a rather small value of 0.15, which simulates a sparsely sampled feature space in the current region of investigation. Since there is almost no interaction between the different examples with respect to this modeling, exploration of almost every part of the space is important and will add valuable information. In contrast to this, the right plot was computed with $\sigma_{\text{RBF}} = 1.5$ which simulates a densely packed feature space. In this case, improving the actual discrimination

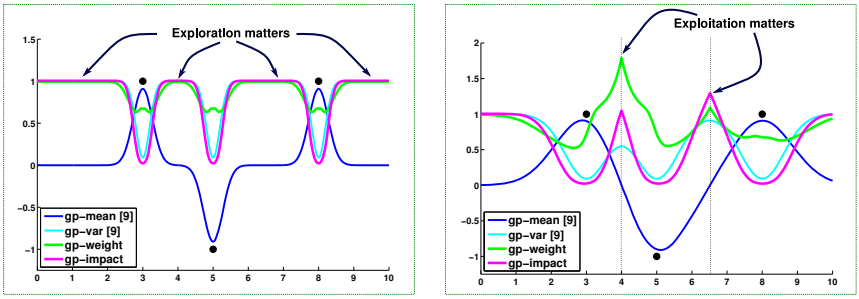

Fig. 2. A synthetic 1D example visualizing the different active learning strategies. In the left plot, known samples indicated by black dots are widely spread (a sparsely sampled feature space), *i.e.*, exploration matters, whereas in the right plot, known samples are close together (a densely packed space) where exploitation is more urgent.

ability of the model is more important and clarifying the actual class boundaries should be in the focus of actively selecting new samples. However, we note that both mean and variance favor samples maximally far away from the current distribution, which leads to outliers that are unrelated to the current problem. We see in the next section that these observations are confirmed in several visual recognition scenarios.

5 Experimental Evaluation

Active learning experiments are conducted on established image categorization datasets. Our findings can be summarized as follows:

1. For a **small number of initial training examples**, our new strategies perform significantly better than state-of-the-art strategies for GP due to a suitable trade-off between exploration and exploitation.
2. For **representative initial training sets**, our new strategies perform as good as established methods and lead to significant performance gains over random sampling.

Experimental Setup. We conduct active learning experiments on popular image recognition datasets. For every dataset, we first randomly select 100 subsets consisting of a single positive and 9 negative classes, and average results over 10 random initializations per subset. In total, this results in 1,000 different active learning scenarios per dataset. Accuracies after every query are evaluated on disjoint test sets using the area under receiver operator curves (AUC). Final learning curves are computed by averaging AUC scores over all subsets and initializations. The noise parameter for model regularization was optimized by maximizing the marginal likelihood. Experiments were conducted in Matlab[2]. Runtimes are not visualized, since the majority of computational time is spent for updating classifier and kernel values. In addition, asymptotic runtimes

[2] Source code is available at `www.inf-cv.uni-jena.de/active_learning`

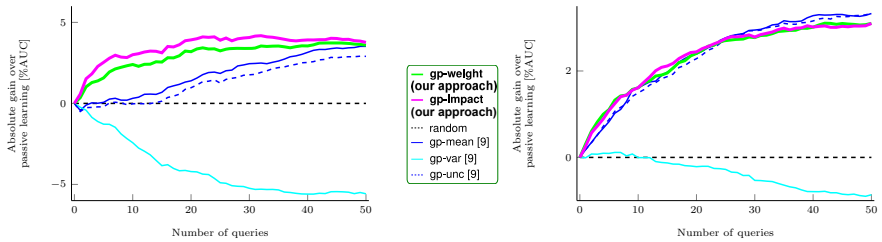

Fig. 3. Gain of active over passive learning on **Caltech-256**. Initial number of samples per class is 1 and 5, respectively.

for the introduced methods are in the same order of complexity as $\mathcal{Q}_{\sigma_*^2}$ and $\mathcal{Q}_{\text{unc}}$. Results are presented for passive learning (random), three state-of-the-art strategies based on a GP model (gp-mean, gp-var, and gp-unc) as well as the two introduced strategies (gp-weight and gp-impact). **Note that absolute performance values are part of the supplementary material.**

Results on Caltech-256. A commonly used benchmark datasets for visual object recognition is Caltech-256, which consists of 256 object categories from different areas of real-world situations, *e.g.*, animals, vehicles, or persons. For the sake of simplicity and reproducibility, we represent images with L_1-normalized bag-of-visual-words histograms over densely sampled SIFT features[3].

In Fig. 3, experimental results averaged over $1,000$ binary settings are visualized[4] . First of all, we observe the prominent gain of active learning methods over passive learning except for variance-based queries. This clearly stresses the benefit of active learning in general. Apart from this, we note that the new strategies outperform existing methods in the presence of few labeled examples and lead to comparable results if more examples are already known. In both settings they outperform random querying by far.

Results on ImageNet. Within the last years, the ImageNet database became a standard benchmark for large-scale visual object classification. In total, $1,000$ different classes from a wide range of different synsets are provided. We again represent images with L_1-normalized bag-of-visual-words histograms over densely sampled SIFT features[5].

In Fig. 4, the experimental results are visualized[4] . If the number of initial samples is low, we again observe that state-of-the-art strategies cannot improve accuracies over passive learning, since the majority of unlabeled samples gives valuable information in this early stage. However, even for this setup our strategies are able to clearly improve results. Especially in medical applications, where obtaining even a handful of labeled samples is extremely expensive, choosing the informative ones is highly valuable. For larger initial sets, we notice that our new methods lead to results comparable to those of state-of-the-art strategies. Again, passive learning is significantly outperformed.

[3] `http://homes.esat.kuleuven.be/~tuytelaa/unsup_features.html`
[4] Further learning curves can be found in the supplementary material.
[5] `http://www.image-net.org/download-features`

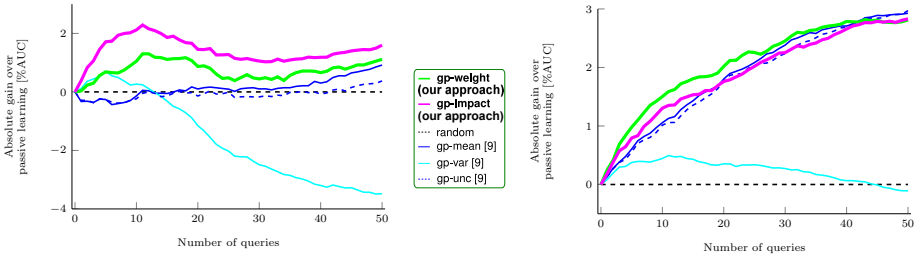

Fig. 4. Gain of active over passive learning on **ImageNet**. Initial number of samples per class is 1 and 5, respectively.

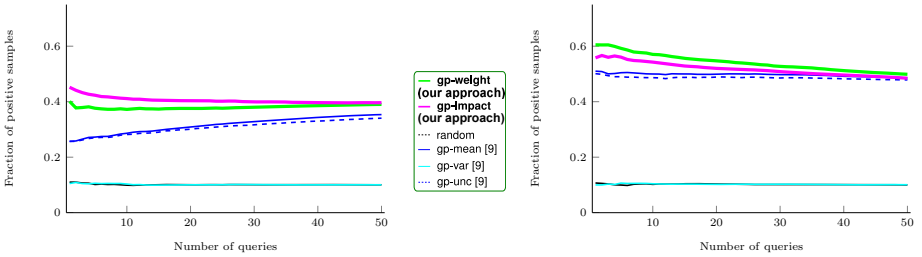

Fig. 5. Fraction of positive examples within queried images averaged over 1,000 experiments on ImageNet. Initial number of samples per class is 1 and 5, respectively.

Fig. 6. Queried images using $\mathcal{Q}_{\text{impact}}$ (*top*) and $\mathcal{Q}_{\sigma_*^2}$ (*bottom*) in a single run on ImageNet. Green and blue borders indicate images of positive and negative classes, respectively. Queries of the remaining strategies can be found in the supplementary material.

One reason for the superior behavior of our strategies can be derived from Fig. 5: both strategies $\mathcal{Q}_{\text{weight}}$ and $\mathcal{Q}_{\text{impact}}$ are able to balance the selection of positive and negative samples, which is essential for building a representative model. In contrast to that, established methods are more biased towards negative examples, which leads to models less useful for the binary classification task. A qualitative example is given in Fig. 6 which further highlights this observation.

6 Conclusion and Further Work

We presented new active learning methods that explicitly select non-redundant (relevant) unlabeled examples for annotation. Our methods are based on measuring the impact of an unlabeled example on a Gaussian process model in the

case the example would be added as a new labeled training example. The resulting active learning approaches were able to outperform previous methods significantly on established benchmark datasets.

Future work will focus on combining active learning methods and learning optimal combinations from auxiliary data, *i.e.*, we would like to perform active learning in a transfer learning fashion, where we learn priors for suitable strategies from other (visual) classification tasks. Another topic is to combine our active learning methods with large-scale Gaussian process approaches [9].

References

1. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. JMLR 5, 255–291 (2004)
2. Bernstein, D.S.: Matrix Mathematics, 2nd edn. Princeton University Press (2009)
3. Bodesheim, P., Rodner, E., Freytag, A., Denzler, J.: Divergence-based one-class classification using gaussian processes. In: BMVC, pp. 50.1–50.11 (2012)
4. Campbell, C., Cristianini, N., Smola, A.: Query learning with large margin classifiers. In: ICML (2000)
5. Cebron, N., Berthold, M.: Active learning for object classification: from exploration to exploitation. Data Mining and Knowledge Discovery 18, 283–299 (2009)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
7. Ebert, S., Fritz, M., Schiele, B.: Ralf: A reinforced active learning formulation for object class recognition. In: CVPR, pp. 3626–3633 (2012)
8. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning 28, 133–168 (1997)
9. 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)
10. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. IJCV 88, 169–188 (2010)
11. Kemmler, M., Rodner, E., Denzler, J.: One-class classification with gaussian processes. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 489–500. Springer, Heidelberg (2011)
12. Osugi, T., Kim, D., Scott, S.: Balancing exploration and exploitation: a new algorithm for active machine learning. In: ICDM, pp. 330–337 (2005)
13. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. The MIT Press (2006)
14. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: ICML, pp. 441–448 (2001)
15. Schohn, G., Cohn, D.: Less is more: Active learning with support vector machines. In: ICML, pp. 839–846 (2000)
16. Seo, S., Wallat, M., Graepel, T., Obermayer, K.: Gaussian process regression: Active data selection and test point rejection. In: IJCNN, pp. 241–246 (2010)
17. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. JMLR 2, 45–66 (2002)
18. Vezhnevets, A., Buhmann, J.M., Ferrari, V.: Active learning for semantic segmentation with expected change. In: CVPR, pp. 3162–3169 (2012)
19. Yeh, T., Darrell, T.: Dynamic visual category learning. In: CVPR, pp. 1–8 (2008)