Active Learning for Regression Tasks with Expected Model Output Changes

Christoph Käding[1]
christoph.kaeding@uni-jena.de
Erik Rodner[2]
erik.rodner@zeiss.com
Alexander Freytag[2]
alexander.freytag@zeiss.com
Oliver Mothes[1]
oliver.mothes@uni-jena.de
Björn Barz[1]
bjoern.barz@uni-jena.de
Joachim Denzler[1]
joachim.denzler@uni-jena.de

[1] Computer Vision Group
Friedrich Schiller University Jena
Jena, Germany

[2] Carl Zeiss AG
Jena, Germany

Abstract

Annotated training data is the enabler for supervised learning. While recording data at large scale is possible in some application domains, collecting reliable annotations is time-consuming, costly, and often a project's bottleneck. Active learning aims at reducing the annotation effort. While this field has been studied extensively for classification tasks, it has received less attention for regression problems although the annotation cost is often even higher. We aim at closing this gap and propose an active learning approach to enable regression applications.

To address continuous outputs, we build on Gaussian process models – an established tool to tackle even non-linear regression problems. For active learning, we extend the expected model output change (EMOC) framework to continuous label spaces and show that the involved marginalizations can be solved in closed-form. This mitigates one of the major drawbacks of the EMOC principle. We empirically analyze our approach in a variety of application scenarios. In summary, we observe that our approach can efficiently guide the annotation process and leads to better models in shorter time and at lower costs.

1 Introduction

As impressive as latest advances in computer vision and machine learning are, the majority of our today's systems crucially depend on the availability of ample annotated data to learn from. Collecting this data can be the limiting factor when building machine learning systems, especially when expert knowledge is needed for reliable annotations. A prominent example are medical diagnosis systems, where reliable annotations should only be provided by a

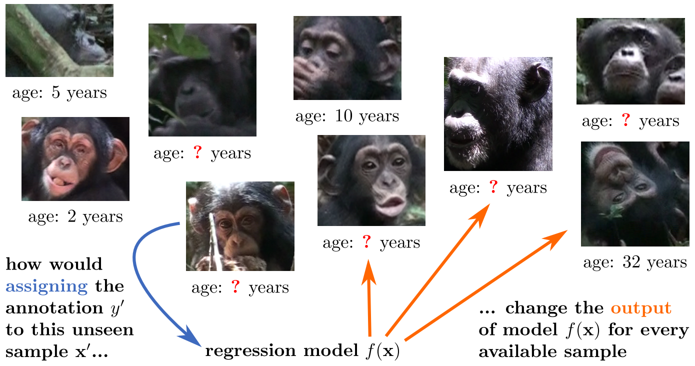
age: 5 years
age: 2 years
age: ? years
age: 10 years
age: ? years
age: ? years
age: ? years
age: ? years
age: 32 years
age: ? years
how would assigning the annotation y' to this unseen sample $\mathbf{x}'$...
regression model $f(\mathbf{x})$
... change the output of model $f(\mathbf{x})$ for every available sample

Figure 1: The idea of the expected model output change criterion for regression tasks: estimate and maximize the change of model outputs via active sample selection.

team of qualified medical practitioners rather than by crowd-annotating systems like Amazon Mechanical Turk [49]. In these scenarios, active learning techniques can reduce the effort of annotating large amounts of data.

Using active learning, an initial (small) annotated set $\mathfrak{L}$ is incrementally extended by carefully choosing examples from a (large) pool of unlabeled data $\mathfrak{U}$ and requesting their annotations. Given a restricted annotation budget, only the most informative examples are selected by maximizing the expected benefit of the newly annotated samples for supervised training. While active learning has been extensively studied for classification tasks (*e.g.,* [37, 57]), it has received less attention for regression problems. We aim at closing this gap since annotation with a continuous output is often even more expensive than for classification.

Our active learning approach for regression is inspired by the expected model output change (EMOC) criterion [14], which is an established active learning technique for classification tasks [37]. By applying EMOC, only those examples are queried that maximize the expected change of model outputs after annotation which serves as a proxy for the expected reduction of errors. A visualization for the application to age regression is shown in Fig. 1.

We transfer EMOC to regression where model outputs are continuous. An exact expectation over these continuous but yet unknown annotations is computationally infeasible without further assumptions. Therefore, we chose a Gaussian process (GP) regression [54] to model the unknown relations between inputs and regression outputs, which has been successfully applied in a variety of application domains (*e.g.,* [8, 9, 18, 36, 54, 65]). This allows not only for efficient training and model updates, but further allows us to apply closed-form solutions to the otherwise intractable marginalization within EMOC.

In a variety of applications and in comparison to state-of-the-art techniques, we show the benefits of our method. We observe that EMOC is consistently able to efficiently guide the annotation process while competitor methods succeed on some data and fail on other.

2 Related Work

Active learning is a widely studied field in machine learning. In the following, we briefly review some influential or related work. A broader overview can be found in [37, 57].

Active Learning for Classification One of the most prominent ideas in active learning is uncertainty sampling, where most uncertain examples with respect to the current model are selected. This uncertainty can be formulated in various ways, for example as distance to a classification boundary [11, 16, 33, 60], or it can be directly derived from the underlying classification model [20, 25, 34]. Other approaches try to combine active learning with other research goals. For example, [21] directly fuses active learning with novelty detection. In [2], the selection of instances is combined with the selection of models for prediction and in [46], even annotators are actively selected. Especially for huge amounts of data, a combination of active learning with hashing is a common technique [26, 28, 29]. Most active learning methods are classifier-independent, but some of them are particularly designed for a single type of prediction model, *e.g.,* CNN architectures [38, 51, 64].

The majority of methods only use surrogate metrics to estimate the future reduction of error. Two exceptions are [61], which describes the information gain of unlabeled samples, and [56], where the value of unlabeled samples is estimated according to the empirical risk. Noteworthy, most of these are not suitable for regression tasks.

Active Learning for Regression Tasks In [7], active learning is combined with passive learning by introducing bounds to the passive learning scheme. The work of [10] describes instance selection for various models mainly based on estimated variance. Exploration guided active learning (EGAL) proposed in [23] is a selection scheme that is based on a combination of diversity and density. An active learning algorithm using Gibbs sampling is proposed in [13]. Here, samples are drawn if estimated labels generated by repeated Gibbs sampling differ. Similar in mind is the work of [59] which presents an approach called ALICE. In contrast to our work, this approach relies on importance-weighted least-squares models and explicitly tries to minimize the generalization error. Active selection based on distances in feature space is presented in [68]. But more importantly, this work discusses the problems active learning has to face for regression tasks and reveals the power of passive learning. Finally, a comparison of different model-based and model-free algorithms is given in [52].

Again, most of the methods presented above are using surrogates and are not clearly linked to risk minimization.

Model and Output Change Our approach relies on expected model output changes, which have been analyzed for different scenarios and model types before. For example, [62] presented an approach to semantic segmentation using decision trees. The works of [14, 30, 31] proposed solutions for classification with GPs. Meanwhile, [5] presented the combination of active and passive learning based on expected changes. Like our approach, these works can be related to risk minimization. However, they are not suitable for regression tasks per se.

Similarly well-studied is the approach of selecting examples that induce large model changes. For example, [6, 24, 58] follow this approach. As shown later, this technique can be expressed as a simplification of our method and is empirically inferior to our approach.

Another related approach is to select samples which are likely to influence the entropy of outputs. A common approach is to minimize the predictive entropy of model outputs and therefore to maximize information gain [57]. In [22], the authors present a formulation of this, called Bayesian active learning by disagreement (BALD), where the goal is to maximize the mutual information between predictions and model posterior. This approach is specially tailored to binary tasks with Gaussian processes involving the assumption that outputs are Bernoulli distributed. Hence, it can not be applied for our regression scenarios as it is.

3 Active Learning for Regression Tasks

In machine learning, we want to estimate a function $f : \Omega \to \mathcal{Y}$ which maps from inputs $x \in \Omega$ to outputs $y \in \mathcal{Y}$. In the original version of EMOC for classification [14], which will be reviewed in the following, the elements of $\mathcal{Y}$ are discrete. We extend it to regression tasks with a continuous label space $\mathcal{Y}$ and show that a closed-form solution of the computationally demanding expectation operation can still be derived.

3.1 Review of EMOC

The main motivation of the EMOC framework is to select the most useful examples for labeling, while avoiding the selection of redundant or irrelevant ones that would not lead to any change of model outputs when added to the training set after annotation. To tackle this goal, EMOC selects examples $x_* \in \mathfrak{U}$ with the largest expected change of model outputs:

$$x_* \; = \; \underset{x' \in \mathfrak{U}}{\operatorname{argmax}} \, \Delta f(x') \; .$$

As shown in [30], EMOC is closely related to the principles of expected error reduction and expected model change (EMC). In contrast to EMC, however, EMOC measures not the distance between old and updated model parameters, but the change of model *outputs* (see also the visualization in Fig. 1):

$$\Delta f(x') \; = \; \mathbb{E}_x \mathbb{E}_{y'|x'} \mathcal{L}\left(f(x), f'(x)\right) \; , \tag{1}$$

with f' being the old model f updated with (x', y'). The maximization of Eq. (1) can also be understood as selecting that sample x' for annotation that *shakes* the current view on the world the most. In the following, we consider general L_P-loss functions, *i.e.*, $\mathcal{L}(f(x), f'(x)) = ||f(x) - f'(x)||_P$. Hence, any suitable L_P-loss function can be selected, in contrast to the fixed L_1-loss, which has been used in [14, 30].

EMOC Framework for Gaussian Process Regression While the EMOC principle can be applied to any machine learning method (*e.g.*, [5, 31, 32, 33, 63]), we follow [14] and use GP regression as a underlying model for f. For a chosen kernel function $\kappa(\cdot,\cdot)$ and a zero-mean assumption, the value predicted by f can be obtained as:

$$f(x') \; = \; k(x')^T \alpha = \sum_{i=1}^{n} \alpha_i \cdot \kappa(X_i, x') \; ,$$

where the vector α results from GP training and represents the weights of each training example $X_i \in \mathfrak{L}$. For the sake of clarity, we use the following abbreviations: $K = \kappa(X, X)$, $k(\cdot) = \kappa(X, \cdot)$, $k'(\cdot) = [k(\cdot), \kappa(x', \cdot)]^T$. The EMOC criterion of Eq. (1) can be rewritten as:

$$\Delta f(x') \; = \; \mathbb{E}_x \underbrace{\mathbb{E}_{y'|x'} \left|\left|k'(x)^T \Delta\alpha\right|\right|_P}_{\doteq \Delta f(x', x)} \; , \tag{2}$$

where $\Delta\alpha$ is the difference between the current model α and the model updated with the labeled example (x', y'). As shown in [14], the model change $\Delta\alpha$ of vanilla GP regression models has a closed-form solution:

$$\Delta\alpha \; = \; \frac{k(x')^T \alpha - y'}{\sigma_n^2 + \sigma_f^2(x')} \begin{bmatrix} \left(K + \sigma_n^2 I\right)^{-1} k(x') \\ -1 \end{bmatrix} \; , \tag{3}$$

where σ_n^2 and $\sigma_f^2(x')$ represent the regularization parameter and the predicted signal variance, respectively. In direct consequence, this allows for closed form evaluations of the EMOC criterion as stated in Eq. (2). However, despite the resulting computational benefits, the criterion still depends linearly on the size of the unlabeled pool due to the expectation operation $\mathbb{E}_x$. To tackle large amounts of unlabeled data, approximation techniques have been presented in [31].

3.2 EMOC for Continuous Label Spaces

In classification, the marginalization of y' can be easily tackled by summing over all possible labels of the discrete output space. Since this is not directly possible for regression problems with $y' \in \mathbb{R}$, we derive a closed form solution for the computation of Eq. (1) for the continuous space $\mathcal{Y}$ of possible labels of a so far unlabeled sample y'.

First, from Eq. (3) it can be observed that the model change $\Delta\alpha$ can be decomposed into a factor depending on y' and a vector $g(x')$ completely independent of the label:

$$\Delta\alpha = g(x') \cdot (k(x')^T \alpha - y') \ .$$

We can now rewrite the expected model output change for the new example x' with respect to a single example x:

$$\begin{aligned}\Delta f(x',x) &= \mathbb{E}_{y'|x'} \left|\left| k'(x)^T g(x')(k(x')^T\alpha - y')\right|\right|_P \\ &= ||\underbrace{k'(x)^T g(x')}_{v}||_P \cdot \mathbb{E}_{y'|x'} ||\underbrace{k(x')^T\alpha}_{c} - y'||_P \ , \end{aligned} \tag{4}$$

with the terms v and c being independent of y'. Substituting $z = y' - c$ leads to:

$$\Delta f(x',x) = ||v||_P \int_{\mathcal{Y}} ||z||_P \, p(z+c\,|\,x')dz \ . \tag{5}$$

In Eq. (5), the posterior distribution of $y' = z + c$ given x' is estimated by the current model. Here, we can exploit that f is modeled as a Gaussian process. Hence, the posterior distribution in Eq. (5) is Gaussian with predictive mean $\mu(x')$ and variance $\sigma_f^2(x')$:

$$p(z+c|x') = \mathcal{N}(z+c\,|\,\mu(x'), \sigma_f^2(x')) \ .$$

This leads to the following expectation operation for z:

$$\begin{aligned}\Delta f(x',x) &= ||v||_P \int ||z||_P \mathcal{N}(z+c|\mu(x'), \sigma_f^2(x'))dz \\ &= ||v||_P \int ||z||_P \mathcal{N}(z|\tilde{\mu}(x'), \sigma_f^2(x'))dz \\ &= ||v||_P \cdot \mathbb{E}[||z||_P] \ , \end{aligned} \tag{6}$$

with $\tilde{\mu}(x') = \mu(x') - c$. Furthermore, Eq. (6) includes the non-central P^{th}-moment of a Gaussian distribution for which a closed-form solution exists involving the confluent hypergeometric function [66]:

$$\mathbb{E}[||z||_P] = \sigma^P \cdot 2^{\frac{P}{2}} \cdot \frac{\Gamma\left(\frac{1+P}{2}\right)}{\sqrt[2]{\pi}} \cdot {}_1F_1\left(-\frac{P}{2}, \frac{1}{2}, -\frac{1}{2}\left(\frac{\tilde{\mu}(x')}{\sigma_f^2(x')}\right)^2\right) \ ,$$

with P being the norm to apply and $\Gamma(\cdot)$ as gamma function. The confluent hypergeometric function ${}_1F_1(\cdot,\cdot,\cdot)$ is defined as follows:

$$ {}_1F_1(a,b,z) \;=\; \sum_{n=0}^{\infty} \frac{a^{(n)} z^n}{b^{(n)} n!} \quad , \tag{7}$$

with $a^{(n)} = a \cdot (a+1) \cdot \ldots \cdot (a+n-1)$ and $a^{(0)} = 1$ ($b^{(n)}$ alike). It can be solved, for example, using the alternative integral definition [1]:

$$ {}_1F_1(a,b,z) \;=\; \frac{\Gamma(b)}{\Gamma(b-a)\Gamma a} \int_0^1 e^{zt} t^{a-1} (1-t)^{b-a-1} dt \quad . $$

However, our proposed criterion can be computationally demanding on very large datasets, *e.g.,* it depends linearly on the amount of all available data $x \in \Omega$ to estimate the model output change for a single sample $x' \in \mathfrak{U}$ (see Eq. (1)). A possibility to overcome this would be to sub-sample the data to approximate $\mathbb{E}_x$ as presented in [31]. Another solution could be to estimate the EMOC score only on the current sample x' itself. It can be shown that this approximation connects our criterion to variance sampling. Please see the supplementary material for further details on this. Another known drawback on large amounts of data arises from the underlying Gaussian process itself and the size of used kernels therein. Therefore, some approaches, as for example [50], are developed to mitigate this.

4 Experiments

Regression is a widely used technique which can be applied in broad range of scenarios. We try to cover as many as possible by conducting experiments in a variety of problem settings including a range of relevant computer vision specific experiments (see Sections 4.1 to 4.4) as well as more general regression problems taken from different machine learning domains (see supplementary material).

Methods We evaluate our method in comparison to several state-of-the-art approaches. In all cases, GP regression is used as underlying model for the regression task to remove additional dependency on the model selection. The simplest baseline is passive learning which is a mere random selection of data points [68] (`random`). A similarly common strategy is uncertainty sampling. Since we use GP models, we are able to compute the predictive variance for any data point, and the sample with the highest predicted variance can be selected [34] (`variance`). Closely related is the selection of samples which maximize the data entropy (`entropy`). The authors of [23] propose a scheme for exploration guided active learning (`EGAL`). Since we are not interested in querying batches of data, we adapt their strategy accordingly. The same authors also introduce a measurement for density and diversity. We use these definitions and maximize the weighted combination of both ($\texttt{Di}\lambda/\texttt{De}1-\lambda$). Another related approach would be to query samples with largest mahalanobis distance in feature space to already labeled samples (`mahalanobis`). The authors of [6] and [58] propose to query samples which are most likely to change the current model (`EMC`). This framework can also be derived from our method as follows. In Eq. (4), we state that the v-term of the EMOC definition consists of the model change and the extended kernel values $k'(x)$. Ignoring these kernel values would lead to a model change criterion which can be defined as follows:

$$ \Delta f(x', x) \;=\; ||g||_P \cdot \mathbb{E}\left[||z||_P \right] \tag{8}$$

	random	variance	EGAL	entropy	Di0.0/De1.0	Di0.5/De0.5	Di1.0/De0.0	mahalanobis	EMC	**EMOC**
AwA	63.78% (2)	69.69% (4)	86.15% (8)	71.51% (7)	87.27% (9)	70.93% (5)	71.23% (6)	67.77% (3)	100.00% (10)	**62.04% (1)**
ABL	59.51% (5)	49.28% (2)	77.17% (8)	77.85% (9)	100.00% (10)	68.85% (6)	69.54% (7)	55.24% (4)	50.92% (3)	**49.24% (1)**
C-Tai	87.00% (2)	91.89% (7)	95.67% (9)	100.00% (10)	94.10% (8)	90.47% (4)	90.58% (5)	91.19% (6)	88.64% (3)	**85.03% (1)**
yearbook	25.49% (3)	33.71% (6)	37.33% (7)	45.16% (9)	100.00% (10)	37.64% (8)	27.87% (4)	23.65% (2)	28.94% (5)	**22.81% (1)**
MSCOCO quality	60.99% (2)	97.01% (8)	61.89% (3)	100.00% (9)	66.27% (6)	64.74% (5)	96.82% (7)	- (10)	63.63% (4)	**60.61% (1)**
average rank	2.80	5.40	7.00	8.80	8.60	5.60	5.80	5.00	5.00	**1.00**

Table 1: Area under error curve in percent relative to the worst performing method on the same dataset (lower is better). Additionally, a ranking (lower is better) of all methods according to their area under error curve per dataset is given in brackets as well as an overall ranking at the bottom.

Evaluation Setup In each experiment, we train an initial model with a common set of labeled data $\mathfrak{L}$. After this, we evaluate active learning by querying single samples out of a pool of yet unknown instances $\mathfrak{U}$. Each newly selected sample is incorporated incrementally into the current model by applying Eq. (3). As a consequence, each of the tested methods starts with the same initially annotated samples, but improves regression over time using newly annotated samples selected by the respective criterion. To assess accuracy, we apply the established root mean square error (RMSE) measure to predictions on the labeled held-out test set $\mathfrak{T}$ at each step. If the variable to predict has more than one dimension, the RMSE score is averaged over the dimensions to yield a single scalar. Hence, we obtain error curves as for example shown in Fig. 2. Due to lack of space, we only present this single error curve in this paper and additionally show all error curves in the supplementary material. Instead, we report the relative improvement in percent of each particular method in Table 1. Additionally, we give the corresponding ranking of the methods for better comparability over different datasets. In our experiments, we follow [14, 30] and set $P = 1$. To compute Eq. (7), we use the function `scipy.special.hyp1f1`(a, b, z) from SciPy [27]. Source code is available at `triton.inf-cv.uni-jena.de/LifelongLearning/gpEMOCreg`.

4.1 Visual Attribute Estimation

Estimating visual attributes from images is a widely studied field. Attributes can be either ordinal or real-valued and can be used for zero-shot learning [42, 43], to categorize object classes utilizing additional text hints [4], or to improve person identification [40, 41]. To demonstrate the effectiveness of our method in those challenging scenarios, we use the established animals with attributes dataset [42] (`AwA`). This dataset consists of 30,475 images from 50 animal classes. Since we are interested in attribute regression instead of attribute based prediction, we use the provided 85 real-valued predicates as output variables.

Experimental Setup For the evaluation, we use an RBF kernel and L2 normalized relu7 features of a VGG19 network provided by the project website [1] which leads to a regression problem with 4,096 input and 85 output dimensions. We conduct three random initializations and use random splits with five initial samples for $\mathfrak{L}$ and 20,000 test samples for $\mathfrak{T}$. After initialization, we perform 1,000 queries on the unlabeled data pool $\mathfrak{U}$ consisting of 10,470 instances. All of these splits are independent of the actual animal class.

Evaluation Results are presented in Fig. 2 as well as in Table 1 and indicate that we are able to perform better than all other evaluated active learning methods. Only the random baseline comes close to our proposed method. A possible reason could be the following:

[1] `http://www.ist.ac.at/~chl/AwA/AwA-features-vgg19.tar.bz2`

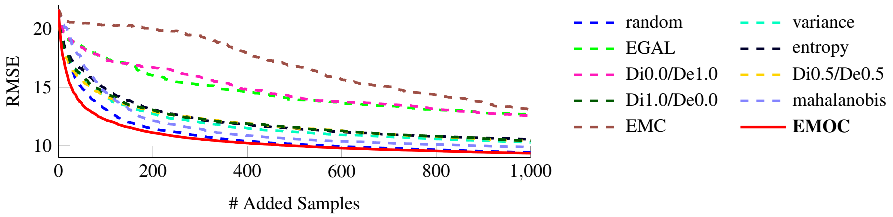

Figure 2: Error curves (lower is better) on the AwA dataset [42].

All active learning methods query samples according to a specific selection scheme which prefers corresponding structures in the feature space. But only passive learning ignores the data structure for selection. Hence, none of the evaluated baselines is able to take advantage of the structure in the data to overcome passive learning. In contrast, results indicate that estimating the change of model outputs makes selection more robust against contrary feature spaces since the importance of feature dimensions is encoded in the model itself.

4.2 Landmark Prediction for Bipedal Locomotion

Biomechanics have strong influence on robotics [12, 53]. Hence, researchers are interested in analyzing animal bipedal locomotion, which requires enormous effort of annotating landmarks (*i.e.,* important anatomical points as joints or bones). For evaluation, we use data provided by zoologists showing quails stepping over obstacles [3]. The walking birds were recorded by a biplanar X-ray acquisition system with a frame rate of 1,000 FPS from two orthogonal views (see supplementary for example images and video material). We conduct an experiment for regression of those landmarks on an already labeled subset of 672 samples (`ABL`). Please note, this approach could also be extended to bounding box regression as for example done by YOLO [55] or facial landmark regression as done by [35].

Experimental Setup We use a deep feature representation from an AlexNet model [39], fine-tuned to distinguish between leg poses of walking birds quantized into ten classes. After fine-tuning, we use activations of conv5 layer concatenated from both views as feature representation for the landmark regression problem. In our evaluation, we consider 15 landmark positions which were normalized to range $[0,1]$. Both lead to a regression problem involving real-valued inputs with 8,192 dimensions and real-valued outputs with 60 dimensions. We average our results over ten experiments and query all 335 unlabeled samples of $\mathfrak{U}$. The performance is measured on random test data $\mathfrak{T}$ comprising also 335 samples. Our GP regression model uses linear kernels and is initialized with two samples for the initial $\mathfrak{L}$.

Evaluation Corresponding results are presented in Table 1. The experiments show that active learning is able to reduce the localization error after already few queries. Our proposed method is superior to mere passive learning and most of the competitor active learning methods. Only variance sampling and EMC obtain similar accuracies which can be attributed to strong relationships between the methods (see Eq. (8) and supplementary material).

4.3 Age Prediction

Age prediction is an important topic in various research areas where domain experts are essential for correct assessment. We apply active learning to age regression on two datasets.

First, we use the C-Tai dataset which was originally introduced by [47, 48] and prepared as benchmark dataset by [15] for computer vision researchers. This dataset consists of images of 5,078 chimpanzees in the wild taken at Tai National Park in Côte d'Ivoire and provides attribute information like age, gender or identity. Since there are missing information for some images, we consider a subset of 4,414 image-age-pairs.

Second, we use the yearbook dataset proposed by [17] which consists of yearbook photos from 115 American high schools from 1905 to 2013. As done by [17], we only consider images of the 20,248 females and try to predict the year the photo was taken.

Experimental Setup For both datasets, we average our results over three random test splits and draw 1,000 samples out of the unlabeled pool $\mathfrak{U}$. We use L2 normalized fc7 features calculated by vanilla AlexNet and RBF kernels. Hence, in both scenarios, regression models with 4,096 real-valued input and a single real-valued output dimension are learned. In terms of the C-Tai dataset, we start with four initial samples for $\mathfrak{L}$ and split the dataset in 2,205 test instances for $\mathfrak{T}$ and 2,205 unlabeled samples in $\mathfrak{U}$. For the yearbook dataset, we initialize our models with three random samples in $\mathfrak{L}$ and use 10,000 samples as unlabeled pool $\mathfrak{U}$ and 10,254 instances as test set $\mathfrak{T}$.

Evaluation The results for the C-Tai and the yearbook datasets are shown in Table 1. Please note that the CNN is not fine-tuned towards age regression. This fact should be considered when explaining the surprisingly good performance of random sampling. We belief that a feature representation adopted towards age regression instead of object recognition will change the results. The development of such an optimized representation is beyond the scope of this paper but subject to future investigations. However, the proposed EMOC framework is able to achieve best results even in these challenging scenarios.

4.4 Image Quality Assessment

A pre-requisite for reliable data analysis by any automated inspection system is the availability of non-degraded data. However, a major source of failure during the life cycle of inspection systems is defect or manipulated hardware which can result in blurry or noisy images. In practice, automated image quality assessment can solve this problem. Similar to [67], we tackle the estimation of image quality as an regression task. In detail, we predict the degree of Gaussian blur and salt-and-pepper noise of potentially disturbed images (MSCOCO quality). In contrast to [67], we directly predict the image quality from extracted CNN features instead of learning a noise type classifier first and an noise-type specific regressor for the noise level thereafter.

Experimental Setup We sample 1,500 random images from the training set of the 2014 MSCOCO v1.0 dataset [45]. After resizing the images to 227 × 227 pixels, a Gaussian blur kernel with randomly selected sigma between zero and five is applied. Finally salt-and-pepper noise is applied to at most 25% of the pixels (some visualizations can be found in the supplementary material). We use a vanilla AlexNet and compute L2 normalized pool2 features as image representation as well as a regression model with RBF kernel. This leads to a regression problem with 43,264 dimensional real-valued inputs and two real-valued outputs. Both, the sigma of the blur kernel and the ratio of salt-and-pepper-noise, is normalized independently to generate the labels. The evaluation protocol uses five initial samples for $\mathfrak{L}$ and 500 samples for the test set $\mathfrak{T}$. All remaining samples serve as unlabeled samples in $\mathfrak{U}$. Presented results are averaged over three random initializations while drawing all remaining samples out of the unlabeled pool $\mathfrak{U}$.

Evaluation Obtained results can be found in Table 1. First of all, results for mahalanobis are missing due to the memory demand of calculations using pool2 features. Since we use an AlexNet which is trained to distinguish between object categories, the obtained features are not designed to estimate image quality degradations. This may result in a feature space where distances between data points are wrong indicators for different blur or noise strength. The poor performance of all methods which rely on pure exploration (*i.e.,* variance, entropy, and diversity) support this intuition. Our proposed EMOC criterion is able to perform best since it is involving the change of model predictions rather than solely relying on those distances.

4.5 Summary of Experimental Results

Some of the previously presented experiments rely on features which were not specifically tailored for regression tasks. To develop adapted representations is behind the scope of the currently presented work. However, to overcome this short-coming, we present additional results in the supplementary material including active sensor placement for wave height estimation on the coastDat-1 dataset [19] as well as experiments on established UCI regression benchmark data [44]. Even on this regression specific data, we are able to show that our proposed EMOC criterion performs best.

Overall, the experiments represent a broad range of application scenarios and are representative for various fields of research. Our evaluations show that the ranking of the investigated methods differs from dataset to dataset. The only exception is our EMOC criterion which is able to achieve the top rank on each dataset (albeit with small margin in few cases). No other method shows a similar consistent behavior over the variety of regression problems. The second-best method for the experiments presented in this paper is random sampling with an average rank of 2.80. In terms of all conducted experiments, including those shown in the supplementary material, the best baseline is diversity sampling with average rank 3.82. In either case, The discrepancy in ranking between EMOC and the strongest competitor method is remarkably large.

Finally, all experiments reveal the surprising strength of passive learning. The counter-intuitive performance of mere random selection of data points for regression tasks was already discussed in [68] and could be proved once more in our evaluation. We conclude that regression tasks are more challenging than classification for active learning. Our intuition is that the real-valued nature of the output space $\mathcal{Y}$ causes this effect, but to the best of our knowledge, a convincing explanation from a learning theory point of view is still missing. A possible consequence could be to combine active with passive learning as suggested by [5].

5 Conclusion

In this paper, we aimed at cutting the annotation costs when training models for regressions problems. Our approach extends the established EMOC criterion from mere classification towards continuous regression tasks. By using Gaussian process regression models, which are a versatile tool for regression applications and offer efficient model updates, we were furthermore able to derive closed-form solutions for the marginalization operations within the EMOC criterion. Since regression tasks arise in many applications, we evaluated our method on a broad range of different datasets. We were able to empirically prove that EMOC leads to the largest error reduction and thereby steers the annotation process most efficiently.

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1972.

[2] Alnur Ali, Rich Caruana, and Ashish Kapoor. Active learning with model selection. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2014.

[3] Emanuel Andrada, Daniel Haase, Yefta Sutedja, John A. Nyakatura, Brandon M. Kilbourne, Joachim Denzler, Martin S. Fischer, and Reinhard Blickhan. Mixed gaits in small avian terrestrial locomotion. *Scientific Reports*, 2015.

[4] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision (ECCV)*, 2010.

[5] Djallel Bouneffouf. Exponentiated gradient exploration for active learning. *Computers*, 2016.

[6] Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *International Conference on Data Mining (ICDM)*, 2013.

[7] Rui Castro, Rebecca Willett, and Robert Nowak. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[8] Tao Chen, Julian Morris, and Elaine Martin. Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems*, 2007.

[9] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research (JAIR)*, 1996.

[11] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: active learning in imbalanced data classification. In *Conference on Information and Knowledge Management*, 2007.

[12] Siyuan Feng, X Xinjilefu, Christopher G Atkeson, and Joohyung Kim. Optimization based controller design and implementation for the atlas robot in the darpa robotics challenge finals. In *International Conference on Humanoid Robots (Humanoids)*, 2015.

[13] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 1997.

[14] Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*, 2014.

[15] Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar Kühl, and Joachim Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition (GCPR)*, 2016.

[16] Chun-Jiang Fu and Yu-Pu Yang. A batch-mode active learning svm method based on semi-supervised clustering. *Intelligent Data Analysis*, 2015.

[17] Shiry Ginosar, Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A Efros. A century of portraits: A visual historical record of american high school yearbooks. In *International Conference on Computer Vision Workshops (ICCV-WS)*, 2015.

[18] Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *International Conference on Machine Learning (ICML)*, 2005.

[19] Zentrum für Material-und Küstenforschung GmbH Helmholtz-Zentrum Geesthacht. coastdat-1 waves north sea wave spectra hindcast (1948-2007), 2012.

[20] Steven CH Hoi, Rong Jin, and Michael R Lyu. Large-scale text categorization by batch mode active learning. In *International Conference on World Wide Web*, 2006.

[21] Timothy M Hospedales, Shaogang Gong, and Tao Xiang. A unifying theory of active discovery and learning. In *European Conference on Computer Vision (ECCV)*, 2012.

[22] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[23] Rong Hu, Sarah Jane Delany, and Brian Mac Namee. Egal: Exploration guided active learning for tcbr. In *International Conference on Case-Based Reasoning*, 2010.

[24] Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active learning for speech recognition: the power of gradients. *arXiv preprint arXiv:1612.03226*, 2016.

[25] Prateek Jain and Ashish Kapoor. Active learning for large multi-class problems. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[26] Prateek Jain, Sudheendra Vijayanarasimhan, and Kristen Grauman. Hashing hyperplane queries to near points with applications to large-scale active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[27] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

[28] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Coverage optimized active learning for k-nn classifiers. In *International Conference on Robotics and Automation (ICRA)*, 2012.

[29] Ajay J Joshi, Fatih Porikli, and Nikolaos P Papanikolopoulos. Scalable active learning for multiclass image classification. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[30] Christoph Käding, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] Christoph Käding, Alexander Freytag, Erik Rodner, Andrea Perino, and Joachim Denzler. Large-scale active learning with approximated expected model output changes. In *German Conference on Pattern Recognition (GCPR)*, 2016.

[32] Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. Active and continuous exploration with deep neural networks and expected model output changes. In *NIPS Workshop on Continual Learning and Deep Networks (NIPS-WS)*, 2016.

[33] Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. Watch, ask, learn, and improve: A lifelong learning cycle for visual recognition. In *European Symposium on Artificial Neural Networks (ESANN)*, 2016.

[34] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision (IJCV)*, 2010.

[35] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[36] Michael Kemmler, Joachim Denzler, Petra Rösch, and Jürgen Popp. Classification of microorganisms via raman spectroscopy using gaussian processes. In *Annual Symposium of the German Association for Pattern Recognition (DAGM)*, 2010.

[37] Adriana Kovashka, Olga Russakovsky, Li Fei-Fei, Kristen Grauman, et al. Crowdsourcing in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 2016.

[38] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. *arXiv preprint arXiv:1511.06789*, 2015.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[40] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. Facetracer: A search engine for large collections of images with faces. In *European Conference on Computer Vision (ECCV)*, 2008.

[41] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision (ICCV)*, 2009.

[42] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[43] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2014.

[44] M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[46] Chengjiang Long, Gang Hua, and Ashish Kapoor. A joint gaussian process model for active visual recognition with expertise estimation in crowdsourcing. *International Journal of Computer Vision (IJCV)*, 2016.

[47] Alexander Loos. Identification of great apes using gabor features and locality preserving projections. In *International Workshop on Multimedia Analysis for Ecological Data*, 2012.

[48] Alexander Loos and Andreas Ernst. An automated chimpanzee identification system using face detection and recognition. *EURASIP Journal on Image and Video Processing*, 2013.

[49] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. Being a turker. In *Conference on Computer Supported Cooperative Work & Social Computing*, 2014.

[50] Pablo Morales-Alvarez, Adrián Pérez-Suay, Rafael Molina, and Gustau Camps-Valls. Remote sensing image classification with large-scale gaussian processes. *Transactions on Geoscience and Remote Sensing*, 2018.

[51] David Novotny, Diane Larlus, and Andrea Vedaldi. I have seen enough: Transferring parts across categories. In *British Machine Vision Conference (BMVC)*, 2016.

[52] Jack O'Neill, Sarah Jane Delany, and Brian MacNamee. Model-free and model-based active learning for regression. In *Advances in Computational Intelligence Systems (NIPS)*, 2017.

[53] Marc Raibert, Kevin Blankespoor, Gabriel Nelson, Rob Playter, and T Bigdog Team. Bigdog, the rough-terrain quadruped robot. In *Proceedings of the 17th World Congress*, 2008.

[54] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[55] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[56] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. In *International Conference on Machine Learning (ICML)*, 2001.

[57] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin, Madison, 2010.

[58] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[59] Masashi Sugiyama. Active learning in approximately linear regression based on conditional expectation of generalization error. *Journal of Machine Learning Research (JMLR)*, 2006.

[60] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research (JMLR)*, 2001.

[61] Deepak Vasisht, Andreas Damianou, Manik Varma, and Ashish Kapoor. Active learning for sparse bayesian multilabel classification. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2014.

[62] A. Vezhnevets, J. M. Buhmann, and V. Ferrari. Active learning for semantic segmentation with expected change. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[63] Alexander Vezhnevets, Joachim M Buhmann, and Vittorio Ferrari. Active learning for semantic segmentation with expected change. In *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[64] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2016.

[65] Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, and Heng Tao Shen. What's wrong with that object? identifying images of unusual objects by modelling the detection score distribution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[66] Andreas Winkelbauer. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.

[67] Ruomei Yan and Ling Shao. Blind image blur estimation via deep learning. *Transactions on Image Processing*, 2016.

[68] Hwanjo Yu and Sungchul Kim. Passive sampling for regression. In *International Conference on Data Mining (ICDM)*, 2010.